Stochastic Dynamic Optimization Under Ambiguity

by

Lauren N. Steimle

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2019

Doctoral Committee:

Professor Brian T. Denton, Chair
Associate Professor Mariel Lavieri
Professor Jon Lee
Professor Susan Murphy, Harvard University
Associate Professor Ambuj Tewari

Lauren N. Steimle
steimle@umich.edu
ORCID iD: 0000-0002-4073-6165

Dedication

For my parents.

Acknowledgements

I would like to thank those who provided their assistance in this research and their support throughout my time working on this dissertation. First and foremost, I would like to express enormous gratitude for my advisor, Dr. Brian Denton. Brian is everything that I could have asked for in a Ph.D. advisor. His guidance and support throughout my doctoral studies has been invaluable to me, and I hope to emulate him as I continue growing as a researcher, teacher, and mentor.

I would also like to thank those who have collaborated with me and provided feedback on the work in this dissertation. I am grateful for the guidance and support from Drs. Rodney Hayward, David Kaufman, Nilay Shah, and Jeremy Sussman. I am also thankful for the support of Vinayak Ahluwalia and Charmee Kamdar who have been an immense joy to work with. I give my appreciation to my committee members, Drs. Mariel Lavieri, Jon Lee, Susan Murphy, and Ambuj Tewari, for their valuable feedback and guidance, as well as the interesting discussions that we have had along the way.

I am grateful for the many mentors that I have had throughout my high school, undergraduate, and graduate careers. I am thankful for my teachers and mentors at Elgin Academy who not only nurtured my love of math and computer science, but also ensured that I would have the writing and communication skills that I would need throughout my higher education in engineering. I am incredibly appreciative of my undergraduate mentor, Dr. Arye Nehorai at Washington University in St. Louis, who encouraged me to pursue a Ph.D. and of Dr. Kathy King who supported me in my pursuit of entering a doctoral program. During my time in graduate school, I have been extremely fortunate to have been surrounded by a supportive group of faculty members. I especially want to thank Drs. Amy Cohn, Mark Daskin, Marina Epelman, Mariel Lavieri, Jon Lee, Monroe Keyserling, Larry Seiford, and Siqian Shen for providing their support, advice, and guidance throughout my graduate studies, as well as their general friendliness.

I would also like to thank the staff and my fellow students in the Industrial and Oper-

ations Engineering department more broadly who have provided a great environment to learn in.

I am also thankful for the friendships that have sustained me throughout the ups and downs of graduate school. I have grown so much personally in the past years because of my friends, and I am forever grateful for those who helped me truly come into my own.

Finally, I would like to thank my family. Their unwavering love and support has meant the world to me.

This work was supported by the National Science Foundation under grant numbers DGE-1256260 (Steimle) and CMMI-1462060 (Denton); any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ACC/AHA American College of Cardiology/ American Heart Association.

B&B branch-and-bound.

B&C branch-and-cut.

BFS best-first search.

BrFS breadth-first search.

CDF cumulative distribution function.

CHD coronary heart disease.

CMDP contextual Markov decision process.

CVD cardiovascular disease.

DFS depth-first search.

DM decision maker.

EVPI expected value of perfect information.

FHS Framingham Heart Study.

FIT fecal immunochemical testing.

HDL high density lipoprotein.

LP linear program.

MDP Markov decision process.

MIP mixed-integer program.

MMDP multi-model Markov decision process.

MVP mean value problem.

PercOpt percentile optimization.

POMDP partially-observable MDP.

QALY quality-adjusted life year.

SBP systolic blood pressure.

SOS special ordered set.

TC total cholesterol.

VSS value of the stochastic solution.

WSU Weight-Select-Update.

WVP weighted value problem.

Abstract

Stochastic dynamic optimization methods are powerful mathematical tools for informing sequential decision-making in environments where the outcomes of decisions are uncertain. For instance, the Markov decision process has found success in many application areas that involve sequential decision-making under uncertainty, including the evaluation and design of treatment and screening protocols for medical decision making. However, the usefulness of these models is only as good as the data used to parameterize them, and multiple competing data sources are common in many application areas, including medicine. Unfortunately, the recommendations that result from the optimization process can be sensitive to the data used and thus, susceptible to the impacts of ambiguity in the choices regarding the model's construction.

To address the issue of ambiguity in Markov decision processes, we introduce the multi-model Markov decision process (MMDP) which generalizes a standard Markov decision process (MDP) by allowing for multiple models of the rewards and transition probabilities. Solution of the MMDP generates a single policy that considers the performance of the policy with respect to the different models of the parameters. This approach allows for the decision maker (DM) to explicitly trade off conflicting sources of data. In this thesis, we study this problem in three parts.

In the first part, we study the *weighted value problem (WVP)* in which the DM's objective is to find a single policy that maximizes the weighted value of expected rewards in each model. We identify two important variants of this problem: the *non-adaptive WVP* in which the DM must specify the decision-making strategy before the outcome of ambiguity is observed and the *adaptive WVP* in which the DM is allowed to adapt to the outcomes of ambiguity. We study the structural properties of these problems and establish important connections to partially-observable MDPs (POMDPs) and stochastic integer programs. To solve these problems, we develop exact methods and fast approximation methods supported by error bounds. Finally, we illustrate the effectiveness and the scalability of our approach

using a case study in preventative blood pressure and cholesterol management that accounts for conflicting published cardiovascular risk models.

In the second part, we leverage the special structure of the non-adaptive WVP to design exact decomposition methods for solving MMDPs with a larger number of models. We present a branch-and-cut (B&C) approach to solve a mixed-integer program (MIP) formulation of the problem and a custom branch-and-bound (B&B) approach. Both approaches leverage the decomposable structure of the problem and allows for the solution of MMDPs with a larger number of models. Numerical experiments show that a customized implementation of B&B significantly outperforms B&C.

In the third part, we extend the MMDP beyond the WVP to consider other objective functions that are sensitive to the ambiguity arising from the existence of multiple models. We summarize existing ambiguity-aware formulations and provide modifications to the B&B procedure to solve these alternate formulations. We compare the solution of the MMDP under these alternative objective functions and compare to a tractable approach for handling ambiguity in the transition probability matrices for MDPs that relies on an assumption that models can be selected independently for different points in the planning horizon. We compare these formulations in two case studies related to MDPs of deteriorating systems. We show that the solution to the mean value problem (MVP), wherein all parameters take on their mean values, can perform quite well with respect to several measures of performance under ambiguity. We also show that a common method for addressing ambiguity can lead to overly aggressive actions. We illustrate that it is possible for this classical approach to perform worse than the policy resulting from the MVP in terms of performance in each MMDP model.

In summary, in this dissertation we present new methods for sequential decision-making under uncertainty in the presence of ambiguity. We consider the problem through the lens of MDPs and stochastic programming and present results for measuring the impact of ambiguity on performance. We analyze alternative forms of the problems, describe the complexity of the problems, develop solution methods, and identify properties of the optimal solutions that provide insight into the effects of ambiguity on optimal policies. Our findings suggest that model averaging may be a suitable approach when the ambiguous parameters are closely concentrated around their mean values. However, in other settings, the impact of ambiguity is much more substantial and our methods outperform the more traditional approaches in these settings. Although we illustrate our methods on decision-making for medical treatment and machine maintenance, the methods we present in this

thesis can be applied to other domains in which optimal sequential decision-making uncertainty is clouded by ambiguity.

Chapter 1

Introduction

Throughout our personal and professional lives, we all must make decisions. Often times, determining the best decisions can be quite complicated because we must weigh the short-term consequences of these decisions against the long-term consequences. Decisions can be further complicated because the future is clouded by uncertainty. While we are able to influence what happens in the future, we are unable to completely control our destiny due to inherent randomness. Although many decision makers (DMs) wish to make decisions that best position themselves to achieve their future goals, these factors can leave DMs unsure how best to proceed. The field of stochastic dynamic optimization describes mathematical tools that can be used to inform decision-making in these challenging settings. Optimization describes the field of mathematics focused on selecting the best set of decisions among the alternatives as measured by an objective function. Dynamic optimization describes the methods used when these decisions are made sequentially over time such that decisions made now may influence the decisions made in the future. Stochastic optimization describes the decision-making setting in which the environment evolves in part due to randomness and in part due to the decisions made. Hence, stochastic dynamic optimization can be succinctly described as the field of sequential decision-making under uncertainty. This field has been quite well-studied and has found success in helping DMs in many areas including inventory management, machine maintenance, finance, and healthcare [10, 56]

While standard stochastic dynamic optimization methods are powerful for informing sequential decision-making under uncertainty, these methods have often ignored another layer of uncertainty that faces DMs: the lack of knowledge around the uncertain environment. That is, we may use mathematical models to represent uncertainty, but often times we do not know the best mathematical model to represent the uncertainty in how a system

evolves over time, which can limit the usefulness of the resulting recommendations. To understand how this limitation can impact decision-making, consider that we are offered a bet that depends on the outcome of a flip of a weighted coin. If we know how the coin is weighted, we can evaluate the expected benefit of each outcome and weigh that outcome with how likely it is to occur. However, if the weight of the coin is unknown, our decision becomes much more difficult as we no longer are certain about what the best mathematical model is to help us guide our decisions. For clarity, throughout this thesis, we will refer to *uncertainty* as the imperfect information about the future which can be characterized via a mathematical model. We refer to *ambiguity* as the imperfect information about the mathematical model itself.

In this thesis, we consider the impact of ambiguity on a particular stochastic dynamic optimization method: the Markov decision process (MDP). The MDP is a mathematical model of sequential decision-making under uncertainty, which models the decision-making process as a controlled stochastic system. The MDP generalizes a Markov chain wherein the DM can take actions to influence the transition dynamics of the system. Standard methods allow for the MDP to be solved quickly and provide the DM with a set of actions that maximize the expected value over the planning horizon. Unfortunately, the optimal course of action, as prescribed by the optimization of the MDP, is sensitive to the probabilities that describe the likelihood of transitions that characterize the stochastic process of the system's progression through the possible states.

In the operations research literature, ambiguity in model parameters is typically handled through one of two paradigms: robust optimization and stochastic optimization. Robust optimization handles ambiguity in the parameters by assuming that the parameters are allowed to vary within an *ambiguity set* (sometimes called an uncertainty set). The typical robust optimization approach is to determine the decisions that will perform the best under the worst-case realization of those parameters when they are allowed to vary within the ambiguity set. Robust optimization has been the standard approach for handling ambiguity in the transition matrices of MDPs. However, the literature has shown that the ambiguity sets require special structure in order to be solved quickly and relaxing this assumption can cause the resulting problems to become computationally intractable.

In this thesis, we present new results about ambiguity in MDPs through a stochastic optimization lens. In stochastic optimization, ambiguous parameters are typically modeled as random variables. In many cases, the ambiguous parameters are modeled as discrete random variables with finite support. When this is the case, the realization of these

parameters can be viewed as possible *scenarios* under which the system might operate. We will consider ambiguity in MDPs by allowing the transition probability and reward parameters of the MDP to be one of a finite set of models.

The theoretical contributions of this thesis were motivated by a specific application of MDPs to cardiovascular disease (CVD) management. The management of CVD is characterized by a series of sequential decisions regarding the best way to treat a patient. If a patient's blood pressure and cholesterol levels are left uncontrolled, the patient is at higher risk of having a serious health event, such as a heart attack or stroke. Therefore, over the course of a patient's adult life, it is suggested that a patient visit their doctor who can observe their blood pressure and cholesterol levels and help the patient make decisions regarding their health. Although lifestyle modifications are typically suggested as the first measure to lower blood pressure and cholesterol, they are frequently ineffective due to challenges associated with maintaining behavioral interventions. Thus, many US adults rely on medications to lower these risk factors which in turn lowers their risk of having a heart attack and stroke. Therefore, it is left to the doctor to make a difficult set of trade-offs. One must weigh the long-term benefit of starting a medication, which lowers a patient's risk of having an adverse health event, with the immediate costs of starting a medication, such as the side effects and monetary costs incurred when taking the medication. However conflicting recommendations that can result from multiple reasonable models of a patient's risk leading to ambiguity in the best course of treatment.

Summary of major contributions. In this thesis, we present methods for sequential decision-making under uncertainty in the presence of ambiguity. We summarize the main contributions from each chapter below.

Chapter 2 presents a new framework for addressing ambiguity in MDPs, which we refer to as the multi-model Markov decision process (MMDP). The main contributions of Chapter 2 are as follows:

- *New Method for Handling Parameter Ambiguity in MDPs.* An MMDP generalizes an MDP to allow for multiple models of the transition probabilities and rewards, each defined on a common state space and action space. In this model formulation, the places a weight on each of the models and seeks to find a single policy that will maximize the weighted value function.

- *Optimal Policies for Two Cases of MMDPs.* It is well-known that for standard MDPs, optimal actions are independent of past realized states and actions; optimal

policies are history independent. We show that, in general, optimal policies for MMDPs may actually be history dependent, making MMDPs more challenging to solve. With the aim of designing policies that are easily translated to practice, we distinguish between two important variants: 1) a case where the DM is limited to policies determined by the current state of the system, which we refer to as the *non-adaptive MMDP*, and 2) a more general case in which the DM attempts to find an optimal history-dependent policy based on all previously observed information, which we refer to as the *adaptive MMDP*.

- *Exact and Approximate Solution Methods.* For medical decision making, the non-adaptive problem is more relevant due to its simplicity and is our primary focus. Unfortunately, the well-known value iteration algorithm for MDPs cannot solve MMDPs to optimality. Fortunately, we are able to formulate a mixed-integer program (MIP) that produces optimal policies. We first test this method on randomly generated problem instances and find that even small instances are difficult to solve; moreover, we find that the differences in objective values between the solutions of the adaptive and the non-adaptive problems are small at best. For larger problem instances though, as one might find in medical decision making applications, models are computationally intractable. Therefore, we introduce a fast approximation algorithm based on backwards recursion that we refer to as the Weight-Select-Update (WSU).

- *Implications for CVD Management.* We establish the effectiveness and scalability of this new modeling approach using a case study that addresses ambiguity in the context of preventive treatment of CVD for patients with type 2 diabetes. Our study demonstrates the ability of MMDPs to blend the information of multiple competing medical studies and directly meet the challenge of designing policies that are easily translated to practice while mitigating the impact of ambiguity that arising from the existence of multiple conflicting models.

Chapter 3 expands upon the exact solution methods for the non-adaptive weighted value problem discussed on Chapter 2, which was shown to be NP-hard. We improve these exact solution methods in Chapter 3. The main contributions of Chapter 3 are:

- *Decomposition Methods.* In this chapter, we present two decomposition methods that leverage the decomposable structure of the problem and allow for the solution of larger MMDPs. We present a branch-and-cut (B&C) algorithm for solving the

MIP formulation of the MMDP presented in Chapter 2, as well as a customized branch-and-bound (B&B) approach which begins by relaxing the requirement that each model of the MMDP must operate under the same policy treats and subsequently adds requirements that the policies must match in certain states of the system and times during the planning horizon.

- *Computational comparison.* We present numerical experiments that compare the time to solve the MMDP using the following three exact solution methods: solving the extensive form of the MIP directly, solving the MIP via B&C, and solving the MMDP using the customized B&B approach. We show that the B&B algorithm outperforms the methods based on the MIP formulation of the MMDP.

- *Numerical study of the impacts of ambiguity.* Because we are able to solve larger MMDPs, we are able to present an analysis of the impact of ambiguity in model parameters on the resulting recommendations from the MMDP. We find that when the models' parameters are concentrated around their mean value, the solution of the mean value problem (MVP), wherein all parameters take on their mean values, provides a near-optimal solution to the weighted value problem in many cases. However, when the models' parameters are distributed further from their mean, there is more benefit to solving the weighted value problem.

Chapter 4 extends the model presented in Chapter 2 to reflect other risk-preferences towards ambiguity represented as a finite number of scenarios, as in the MMDP.

- *B&B for alternative risk preferences.* We compile recent advances in MMDPs that consider alternative risk preferences towards ambiguity. We show that these formulations are also solved with minor modifications to the B&B procedure presented in Chapter 3.

- *Numerical study on the mitigation of ambiguity.* The flexibility of the B&B procedure to incorporate other risk preferences and its success in solving moderately-sized MMDPs allows us to perform one of the first analyses comparing various proposed methods in terms of their effectiveness in mitigating the impact of ambiguity on finite-horizon MDPs. We compare alternative formulations of the MMDP and evaluate the resulting policies in terms of their performance on several metrics. These alternative formulations show that the MVP does well on a variety of metrics for

finite-horizon MDPs. We also show that the DM should use caution when using methods described by earlier work if the assumptions required do not hold, as these can produce policies that perform worse than simply using the MVP's policy.

The novel contributions of this thesis are embodied in Chapters 2-4, which present the findings described above. The thesis concludes with Chapter 5 which presents a summary of the most important findings and an outline of opportunities for future research that stem from this work.

Chapter 2

Multi-model Markov Decision Processes

2.1. Introduction

The MDP is a mathematical framework for sequential decision making under uncertainty that has informed decision making in a variety of application areas including inventory control, scheduling, finance, and medicine [10, 56]. MDPs generalize Markov chains in that a DM can take actions to influence the rewards and transition dynamics of the system. When the transition dynamics and rewards are known with certainty, standard dynamic programming methods can be used to find an optimal policy, or set of decisions, that will maximize the expected rewards over the planning horizon.

Unfortunately, the estimates of rewards and transition dynamics used to parameterize the MDPs are often imprecise and lead the DM to make decisions that do not perform well with respect to the true system. The imprecision in the estimates arises because these values are typically obtained from observational data or from multiple external sources. When the policy found via an optimization process using the estimates is evaluated under the true parameters, the performance can be worse than anticipated [46]. This motivates the need for MDPs that account for this ambiguity in the MDP parameters.

In this chapter, we are motivated by situations in which the DM relies on external sources to parameterize the model but has multiple credible choices which provide potentially conflicting estimates of the parameters. In this situation, the DM may be grappling with the following questions: Which source should be used to parameterize the model? What are the potential implications of using one source over another? To address these questions, we propose a new method that allows the DM to simultaneously consider multiple models of the MDP parameters and create a policy that balances the performance while being no

more complicated than an optimal policy for an MDP that only considers one model of the parameters.

2.1.1. Applications to medical decision making

We are motivated by medical applications for which Markov chains are among the most commonly used stochastic models for decision making. A keyword search of the US Library of Medicine Database using PubMed from 2007 to 2017 reveals more than 7,500 articles on the topic of Markov chains. Generalizing Markov chains to include decisions and rewards, MDPs are useful for designing optimal treatment and screening protocols, and have found success doing so for a number of important diseases; e.g., end-stage liver disease [2], HIV [64], breast cancer [4], and diabetes [47].

Despite the potential of MDPs to inform medical decision making, the utility of these models is often at the mercy of the data available to parameterize the models. The transition dynamics in medical decision making models are often parameterized using longitudinal observational patient data and/or results from the medical literature. However, longitudinal data are often limited due to the cost of acquisition, and therefore transition probability estimates are subject to statistical uncertainty. Challenges also arise in controlling observational patient data for bias and often there are unsettled conflicts in the results from different clinical studies; see Mount Hood 4 Modeling Group [51], Etzioni et al. [22], and Mandelblatt et al. [44] for examples in the contexts of breast cancer, prostate cancer, and diabetes, respectively.

A specific example, and one that we will explore in detail, is in the context of CVD for which cardiovascular risk calculators estimate the probability of a major cardiovascular event, such as a heart attack or stroke. There are multiple well-established risk calculators in the clinical literature that could be used to estimate these transition probabilities, including the American College of Cardiology/ American Heart Association (ACC/AHA) Risk Estimator [27] and the risk equations resulting from the Framingham Heart Study (FHS) [75, 76]. However, these two credible models give conflicting estimates of a patient's risk of having a major cardiovascular event. Steimle and Denton [69] showed that the best treatment protocol for CVD is sensitive to which of these conflicting estimates are used leaving an open question as to which clinical study should be used to parameterize the model.

The general problem of multiple conflicting models in medical decision making has also

been recognized by others (in particular, Bertsimas, Silberholz, and Trikalinos [8]), but it has not been addressed previously in the context of MDPs. As pointed out in a report from the Cancer Intervention and Surveillance Modeling Network regarding a comparative modeling effort for breast cancer, the authors note that "the challenge for reporting multimodel results to policymakers is to keep it (nearly) as simple as reporting one-model results, but with the understanding that it is more informative and more credible. We have not yet met this challenge" [31]. This highlights the goal of designing policies that are as easily translated to practice as those that optimize with respect to a single model, but with the robustness of policies that consider multiple models. The primary contribution of our work is meeting this challenge for MDPs.

The general problem of coping with multiple (potentially valid) choices of data for medical decision making motivates the following more general research questions: How can we improve stochastic dynamic optimization methods to account for parameter ambiguity in MDPs? Further, how much benefit is there to mitigating the effects of ambiguity?

2.1.2. Contributions

In this chapter, we present a new approach for handling parameter ambiguity in MDPs, which we refer to as the MMDP. An MMDP generalizes an MDP to allow for multiple models of the transition probabilities and rewards, each defined on a common state space and action space. In this model formulation, the DM places a weight on each of the models and seeks to find a single policy that will maximize the weighted value function. This model was proposed concurrently by Steimle, Kaufman, and Denton [70] for finite-horizon MDPs and by Buchholz and Scheftelowitsch [14] for infinite-horizon MDPs under the name of *concurrent MDPs*.

It is well-known that for standard MDPs, optimal actions are independent of past realized states and actions; optimal policies are history independent. We show that, in general, optimal policies for MMDPs may actually be history dependent, making MMDPs more challenging to solve. With the aim of designing policies that are easily translated to practice, we distinguish between two important variants: 1) a case where the DM is limited to policies determined by the current state of the system, which we refer to as the non-adaptive MMDP, and 2) a more general case in which the DM attempts to find an optimal history-dependent policy based on all previously observed information, which we refer to as the MMDP. We show that the adaptive problem is a special case of a partially-observable

MDP (POMDP) that is PSPACE-hard, and we show that the non-adaptive problem is NP-hard.

For medical decision making, the non-adaptive problem is more relevant due to its simplicity and is our primary focus. Unfortunately, the well-known value iteration algorithm for MDPs cannot solve MMDPs to optimality. Fortunately, we are able to formulate a MIP that produces optimal policies. We first test this method on randomly generated problem instances and find that even small instances are difficult to solve; moreover, we find that the differences in objective values between the solutions of the adaptive and the non-adaptive problems are small at best. For larger problem instances though, as one might find in medical decision making applications, models are computationally intractable. Therefore, we introduce a fast approximation algorithm based on backward recursion that we refer to as the WSU.

Finally, we establish the effectiveness and scalability of this new modeling approach using a case study that addresses ambiguity in the context of preventive treatment of CVD for patients with type 2 diabetes. Our study demonstrates the ability of MMDPs to blend the information of multiple competing medical studies (ACC/AHA and FHS) and directly meet the challenge of designing policies that are easily translated to practice while being robust to ambiguity arising from the existence of multiple conflicting models.

2.1.3. Organization of the chapter

The remainder of this chapter is organized as follows: In Section 2.2, we provide some important background on MDPs and discuss the literature that is most related to our work. We formally define the MMDP in Section 2.3, and in Section 2.4 we present analysis of our proposed MMDP model. In Section 2.5, we discuss exact solution methods as well as fast and scalable approximation methods that exploit the model structure. We test these approximation algorithms on randomly generated problem instances and describe the results in Section 2.6. In Section 2.7, we present our case study. Finally, in Section 2.8, we summarize the most important findings from our research and discuss the limitations and opportunities for future research.

2.2. Background

In this chapter, we focus on discrete-time, finite-horizon MDPs under ambiguity. In this section, we will describe the MDP and parameter ambiguity, as well as the related work aimed at mitigating the effects of ambiguity in MDPs.

2.2.1. Markov decision processes

MDPs are a common framework for modeling sequential decision-making that influences a stochastic reward process. For ease of explanation, we introduce the MDP as an interaction between an exogenous actor, *nature*, and the DM. The sequence of events that define the MDP are as follows: first, nature randomly selects an initial state $s_1 \in \mathcal{S}$ according to the initial distribution $\mu_1 \in \mathscr{M}(\mathcal{S})$, where $\mathscr{M}(\cdot)$ denotes the set of probability measures on the discrete set. The DM observes the state $s_1 \in \mathcal{S}$ and selects an action $a_1 \in \mathcal{A}$. Then, the DM receives a reward $r_1(s_1, a_1) \in \mathbb{R}$ and then nature selects a new state $s_2 \in \mathcal{S}$ with probability $p_1(s_2 \mid s_1, a_1) \in [0, 1]$. This process continues whereby for any decision epoch $t \in \mathcal{T} \equiv \{1, \ldots, T\}$, the DM observes the state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$, and receives a reward $r_t(s_t, s_t)$, and nature selects a new state $s_{t+1} \in \mathcal{S}$ with probability $p_t(s_{t+1} \mid s_t, a_t)$. The DM selects the last action at time T which may influence which state is observed at time $T + 1$ through the transition probabilities. Upon reaching $s_{T+1} \in \mathcal{S}$ at time $T + 1$, the DM receives a terminal reward of $r_{T+1}(s_{T+1}) \in \mathbb{R}$. Future rewards are discounted at a rate of $\alpha \in (0, 1]$ which accounts for the preference of rewards received now over rewards received in the future. In this chapter, we assume without loss of generality that the discount factor is already incorporated into the reward definition. We will refer to the times at which the DM selects an action as the set of *decision epochs*, $\mathcal{T}$, the set of rewards as $R \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{T}|}$, and the set of transition probabilities as $P \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times\mathcal{T}|}$ with elements satisfying $p_t(s_{t+1} \mid s_t, a_t) \in [0, 1]$ and $\sum_{s_{t+1}\in\mathcal{S}} p_t(s_{t+1} \mid s_t, a_t) = 1,\ \forall t \in \mathcal{T},\ s_t \in \mathcal{S},\ a_t \in \mathcal{A}$. Throughout the remainder of this chapter, we will use the tuple $(\mathcal{T}, \mathcal{S}, \mathcal{A}, R, P, \mu_1)$ to summarize the parameters of an MDP.

The realized value of the DM's sequence of actions is the total reward over the planning horizon:

$$\sum_{t=1}^{T} r_t(s_t, a_t) + r_{T+1}(s_{T+1}). \tag{2.1}$$

The objective of the DM is to select the sequence of actions such that the expectation of

(2.1) with respect to the distribution defined by the transition probabilities is maximized. Thus, the DM will select the actions at each decision epoch based on some information available to her. The strategy by which the DM selects the action for each state at decision epoch $t \in \mathcal{T}$ is called a *decision rule*, $\pi_t \in \Pi_t$, and the set of decision rules over the planning horizon is called a *policy*, $\pi \in \Pi$.

There exist two dichotomies in the classes of policies that a DM may select from: 1) history-dependent vs. Markov, and 2) randomized vs. deterministic. History-dependent policies may consider the entire history of the MDP, $h_t := (s_1,\ a_1,\ \ldots,\ a_{t-1},\ s_t)$, when prescribing which action to select at decision epoch $t \in \mathcal{T}$, while Markov policies only consider the current state $s_t \in \mathcal{S}$ when selecting an action. Randomized policies specify a probability distribution over the action set, $\pi_t(s_t) \in \mathscr{M}(\mathcal{A})$, such that action $a_t \in \mathcal{A}$ will be selected with probability $\pi_t(a_t|s_t)$. Deterministic policies specify a single action to be selected with probability 1. Markov policies are a subset of history-dependent policies, and deterministic policies are a subset of randomized policies. For standard MDPs, there is guaranteed to be a Markov deterministic policy that maximizes the expectation of (2.1) [Proposition 4.4.3 in 56] which allows for efficient solution methods that limit the search for optimal policies to the Markov deterministic (MD) policy class, $\pi \in \Pi^{MD}$. We will distinguish between history-dependent (H) and Markov (M), as well as randomized (R) and deterministic (D), using superscripts on Π. For example, Π^{MR} denotes the class of Markov randomized policies.

To summarize, given an MDP $(\mathcal{T}, \mathcal{S}, \mathcal{A}, R, P, \mu_1)$, the DM seeks to find a policy π that maximizes the expected rewards over the planning horizon:

$$\max_{\pi \in \Pi} \mathbb{E}^{\pi, P, \mu_1} \left[\sum_{t=1}^{T} r_t(s_t, a_t) + r_{T+1}(s_{T+1}) \right]. \tag{2.2}$$

A standard MDP solution can be computed in polynomial time because the problem decomposes when the search over Π is limited to the Markov deterministic policy class, Π^{MD}. We will show that this and other properties of MDPs no longer hold when parameter ambiguity is considered.

2.2.2. Parameter ambiguity and related work

MDPs are known as models of sequential decision making under uncertainty. However, this "uncertainty" refers to the imperfect information about the future state of the system

after an action has been taken due to stochasticity. The transition probability parameters are used to characterize the likelihood of these future events. For the reasons described in Section 2.1, the model parameters themselves may not be known with certainty. As a reminder, we will refer to *uncertainty* as the imperfect information about the future which can be characterized via a set of transition probability parameters. We refer to *ambiguity* as the imperfect information about the transition probability parameters themselves.

In this chapter, we consider a variation on MDP in which parameter ambiguity is expressed through multiple models of the underlying Markov chain, and the goal of the DM is to find a policy that maximizes the weighted performance across these different models. The concept of multiple models of parameters is seen in the stochastic programming literature whereby each set corresponds to a "scenario" representing a different possibility for the problem data [9]. Stochastic programming problems typically consist of multiple stages during which the DM has differing levels of information about the model parameters. For example, in a two-stage stochastic program, the DM selects initial actions during the first-stage before knowing which of the multiple scenarios will occur. The DM subsequently observes which scenario is realized and takes *recourse* actions in the second stage. In contrast, in MMDP, the model parameters will never be revealed to the DM.

Perhaps the most closely related research to this chapter is that of Bertsimas, Silberholz, and Trikalinos [8] who recently addressed ambiguity in simulation modeling in the context of prostate cancer screening. The authors propose solving a series of optimization problems via an iterated local search heuristic to find screening protocols that generate a Pareto optimal frontier on the dimensions of average-case and worst-case performance in a set of different simulation models. This article identified the general problem of multiple models in medical decision making; however, they do not consider this issue in MDPs. The concept of multiple models of problem parameters in MDPs has mostly been used as a form of sensitivity analysis. For example, Craig and Sendi [17] propose bootstrapping as a way to generate multiple sets of problem parameters under which to evaluate the robustness of a policy to variation in the transition probabilities. There has been less focus on finding policies that perform well with respect to multiple models of the problem parameters in MDPs. As pointed out in a report from the Cancer Intervention and Surveillance Modeling Network regarding a comparative modeling effort for breast cancer, the authors note that "the challenge for reporting multi-model results to policymakers is to keep it (nearly) as simple as reporting one-model results, but with an understanding that it is more informative and more credible. We have not yet met this challenge" [31]. This

highlights the goal of designing policies that are as easily translated to practice as those that optimize with respect to a single model, but with the robustness of policies that consider performance in multiple models.

The approach of incorporating multiple models of parameters is also seen in the reinforcement learning literature, however the objective of the DM in these problems is different than the objective of the DM in this chapter. For example, consider what is perhaps the most closely related reinforcement learning problem: the contextual Markov decision process (CMDP) proposed by Hallak, Di Castro, and Mannor. The CMDP is essentially the same as the MMDP set-up in that one can think of the CMDP as C MDPs all defined on the same state space and action space, but with different reward and transition probability parameters. In the CMDP problem, the DM will interact with the CMDP throughout a series of episodes occurring serially in time. At the beginning of the interaction, the DM neither has any information about any of the C MDP's parameters, nor does she know which MDP she is interacting with at the beginning of each episode. Our work differs from that of [32] in that we assume the DM has a complete characterization of each of the MDPs, but due to ambiguity the DM still does not know which MDP she is interacting with. Others have studied related problems in the setting of *multi-task reinforcement learning* [13]. Our work differs from this line of research in that we are motivated by problems with shorter horizons while contextual and multi-task learning is appropriate for problems in which the planning horizon is sufficiently long to observe convergence of estimates to their true parameters based on a dynamic learning process, such as in the area of mobile health [33, 50].

Our research is distinct from the more traditional approach of mitigating parameter ambiguity in MDPs, known as *robust dynamic programming*, which represents parameter ambiguity through an ambiguity set formulation. The standard robust dynamic programming is a "max-min" approach in which the DM seeks to find a policy that maximizes the worst case performance when the transition probabilities are allowed to vary within an ambiguity set. The ambiguity set can be constructed as intervals around a point estimate, and the max-min approach represents that the DM is risk neutral with respect to uncertainty and risk adverse with respect to ambiguity.

A key result regarding the max-min problem is that it is tractable for instances that satisfy the (s, a)-*rectangularity* property [35, 52]. The (s, a)-rectangularity implies that observing the realization of a transition probability parameter gives no information about the values of other parameters for any other state-action-time triplet. Because each param-

eter value for any given state-action-time triplet is independent of the others, the problem can be decomposed so that each worst-case parameter is found via an optimization problem called the *inner problem*. Iyengar [35] and Nilim and El Ghaoui [52] provide algorithms for solving the max-min problem for a variety of ambiguity sets by providing polynomial-time methods for solving the corresponding inner problem.

While (s, a)-rectangular ambiguity sets are desirable from a computational perspective, they can give rise to policies that are overly-conservative because the DM must account for the possibility that parameters for each state-action-time triplet will take on their worst-case values simultaneously. Therefore, much of the research in robust dynamic programming has focused on ways to mitigate the effects of parameter ambiguity while avoiding policies that are overly conservative by either finding non-(s, a)-rectangular ambiguity sets that are tractable for the max-min problem or optimizing with respect to another objective function usually assuming some *a priori* information about the model parameter [18, 30, 41, 45, 62, 74, 78].

To our knowledge, Le Tallec [40], Ahmed et al. [1], and Merakli [49], and Saghafian [59] are the only articles that have considered addressing ambiguity in the MDP parameters by using multiple discrete sets of parameters. Le Tallec introduced the concept of an MDP with "random uncertainty" wherein ambiguity is represented as a finite number of models. The author does so as a way to study the complexity of MDPs with ambiguity, but the focus is limited primarily to the complexity of such problems rather than the solution of such problems. Recently, Ahmed et al. propose sampling rewards and transition probabilities at each time step to generate a set of discrete MDPs and then seek to find one policy that minimizes the maximum regret over the set of MDPs. To do this, they formulate a MIP to approximate an optimization problem with quadratic constraints which minimizes regret. They also propose cumulative expected myopic regret as a measure of regret for which dynamic programming algorithms can be used to generate an optimal policy. The authors require that the sampled transition probabilities and rewards are stage-wise independent, satisfying the (s, a)-rectangularity property. Concurrently with our work, Merakli propose percentile optimization approach for MDPs where ambiguity is represented using a finite number of models. In the POMDP setting, Saghafian uses multiple models of the parameters to address ambiguity in transitions among the core states in a POMDP and use an objective function that weights the best-case and worst-case value-to-go across the models. This is in contrast to our work which considers the expected value-to-go among multiple models. Saghafian assumes that the best-case and worst-

case model are selected independently across decision epochs, satisfying the rectangularity assumption. In our MMDP formulation, the rectangularity assumption is not required; the objective is to find a single policy that will perform well in each of the models which may have interdependent transition probabilities across the planning horizon.

Later in this article, we will describe a case study that illustrates the effectiveness and scalability of the MMDP formulation on a medical decision making problem with parameter ambiguity in the context of prevention of cardiovascular disease. As pointed out in Section 2.1, MDPs are increasingly used for designing optimal treatment and screening protocols; however, the literature on addressing ambiguity in MDPs for medical decision making is very sparse. As mentioned previously, Bertsimas, Silberholz, and Trikalinos [8] addressed ambiguity in simulation modeling in the context of prostate cancer screening. Goh et al. [28] proposed finding the best-case and worst-case transition probability parameters for this policy when these parameters are allowed to vary within an ambiguity set. The authors assumed that this ambiguity set is a row-wise independent set that generalizes the existing row-wise ambiguity models in Iyengar [35] as well as Nilim and El Ghaoui [52]. This rectangularity assumption allows for the authors to solve a semi-infinite linear program (LP) problem efficiently. The authors apply their methods to fecal immunochemical testing (FIT) for colorectal cancer and show that, despite the ambiguity in model parameters related to FIT, this screening tool is still cost-effective relative to the most prevalent method, colonoscopy.

To our knowledge, the optimal design of medical screening and treatment protocols under parameter ambiguity is limited to the work of Kaufman, Schaefer, and Roberts [37], Sinha, Kotas, and Ghate [66], and Zhang, Steimle, and Denton [80]. Kaufman, Schaefer, and Roberts [37] consider the optimal timing of living-donor liver transplantation, for which some critical health states are seldom visited historically. They use the robust MDP framework, modeling ambiguity sets as confidence regions based on relative entropy bounds. The resulting robust solutions are of a simple control-limit form that suggest transplanting sooner, when patients are healthier, than otherwise suggested by traditional MDP solutions based on maximum likelihood estimates of transition probabilities. Sinha, Kotas, and Ghate [66] use a robust MDP formulation for response-guided dosing decisions in which the dose-response parameter is allowed to vary within an interval ambiguity set and show that a monotone dosing policy is optimal for the robust MDP. Zhang, Steimle, and Denton [80] propose a robust MDP framework in which transition probabilities are confined to statistical confidence intervals. They employ a rectangularity assumption implying inde-

pendence of rows in the transition probability matrix. They assume an adversarial model in which the DM decides on a policy, and an adversary optimizes the choice of transition probabilities that minimizes expected rewards subject to an "uncertainty budget" on the choice of transition probabilities. While these articles address parameter ambiguity in the transition probabilities, they all assume an (s, a)-rectangular ambiguity set which decouples the ambiguity across decision epochs and states. In contrast, the MMDP formulation that we propose allows for the ambiguity in model parameters to be linked across tuples of states, actions, and decision epochs.

2.3. Multi-model Markov decision processes

In this section, we introduce the detailed mathematical formulation of the MMDP starting with the following definition:

Definition 2.1 (Multi-model Markov decision process)**.** *An MMDP is a tuple $(\mathcal{T}, \mathcal{S}, \mathcal{A}, \mathcal{M}, \Lambda)$ where $\mathcal{T}$ is the set of decision epochs, $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces respectively, $\mathcal{M}$ is the finite discrete set of models, and $\Lambda := \{\lambda_1, \ldots, \lambda_{|\mathcal{M}|}\}$ is the set of exogenous models weights with $\lambda_m \in (0, 1), \forall m \in \mathcal{M}$ and $\sum_{m \in \mathcal{M}} \lambda_m = 1$. Each model $m \in \mathcal{M}$ is an MDP, $(\mathcal{T}, \mathcal{S}, \mathcal{A}, R^m, P^m, \mu_1^m)$, with a unique combination of rewards, transition probabilities, and initial distribution.*

The requirement that $\lambda_m \in (0, 1)$ is to avoid the trivial cases: If there exists a model $m \in \mathcal{M}$ such that $\lambda_m = 1$, the MMDP would reduce to a standard MDP. If there exists a model $m \in \mathcal{M}$ such that $\lambda_m = 0$, then the MMDP would reduce to an MMDP with a smaller set of models, $\mathcal{M} \setminus \{m\}$. The model weights, Λ, may be selected via expert judgment to stress the relative importance of each model, as tunable parameters which the DM can vary (as illustrated in the case study in Section 2.7), according to a probability distribution over the models, or as uninformed priors when each model is considered equally reputable (as in [8]).

In an MMDP, the DM considers the expected rewards of the specified policy in the multiple models. The value of a policy $\pi \in \Pi$ in model $m \in \mathcal{M}$ is given by its expected rewards evaluated with model m's parameters:

$$v^m(\pi) := \mathbb{E}^{\pi, P^m, \mu_1^m} \left[\sum_{t=1}^{T} r_t^m(s_t, a_t) + r_{T+1}^m(s_{T+1}) \right].$$

We associate any policy, $\pi \in \Pi$, for the MMDP with its *weighted value*:

$$W(\pi) := \sum_{m \in \mathcal{M}} \lambda_m v^m(\pi) = \sum_{m \in \mathcal{M}} \lambda_m \mathbb{E}^{\pi, P^m, \mu_1^m} \left[\sum_{t=1}^{T} r_t^m(s_t, a_t) + r_{T+1}^m(s_{T+1}) \right]. \tag{2.3}$$

Thus, we consider the *weighted value problem (WVP)* in which the goal of the DM is to find the policy $\pi \in \Pi$ that maximizes the weighted value defined in (2.3):

Definition 2.2 (Weighted value problem)**.** *Given an MMDP* $(\mathcal{T}, \mathcal{S}, \mathcal{A}, \mathcal{M}, \Lambda)$*, the WVP is defined as the problem of finding a solution to:*

$$W^* := \max_{\pi \in \Pi} W(\pi) = \max_{\pi \in \Pi} \left\{ \sum_{m \in \mathcal{M}} \lambda_m \mathbb{E}^{\pi, P^m, \mu_1^m} \left[\sum_{t=1}^{T} r_t^m(s_t, a_t) + r_{T+1}^m(s_{T+1}) \right] \right\} \tag{2.4}$$

and a set of policies $\Pi^* := \{\pi^* \ : \ W(\pi^*) = W^*\} \subseteq \Pi$ *that achieve the maximum in* (2.4)*.*

The WVP can be viewed as an interaction between the DM (who seeks to maximize the expected weighted value of the MMDP) and nature. In many robust formulations, nature is viewed as an adversary which represents the risk-aversion to ambiguity in model parameters. However, in the WVP, nature plays the role of a neutral counterpart to the DM. In this interaction, the DM knows the complete characterization of each of the models, and nature selects which model will be given to the DM by randomly sampling according to the probability distribution defined by $\Lambda \in \mathscr{M}(\mathcal{M})$. For a fixed model $m \in \mathcal{M}$, there will exist an optimal policy for m that is Markov (i.e., $\pi_m^* \in \Pi^M$). We will focus on the problem of finding a policy that achieves the maximum in (2.4) when $\Pi = \Pi^M$. We will refer to this problem as the *non-adaptive problem* because we are enforcing that the DM's policy be based solely on the current state, and she cannot adjust her strategy based on what sequences of states she has observed. As we will show, unlike traditional MDPs, the restriction to Π^M may not lead to an overall optimal solution. For completeness, we will also describe an extension, called the *adaptive problem*, where the DM can utilize information about the history of observed states, however this extension is not the primary focus of this article. The evaluation of a given policy in the WVP is illustrated in Figure 2.1.

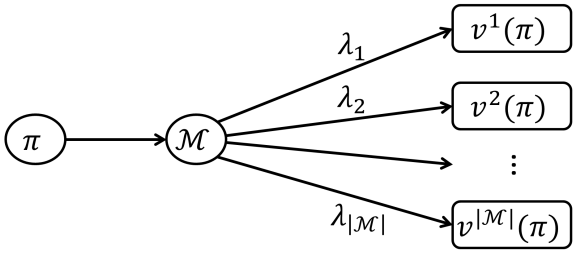


Figure 2.1: An illustration of policy evaluation in terms of weighted value, which is the objective function used to compare policies for an MMDP. The DM specifies a policy π which is subsequently evaluated in each of the $|\mathcal{M}|$ models. The weighted value of a policy π is determined by taking of the sum of this policy's value in each model m, $v^m(\pi)$, weighted by the corresponding model weight λ_m.

2.3.1. The non-adaptive problem

The *non-adaptive problem* for MDPs is an interaction between nature and the DM. In this interaction, the DM specifies a Markov policy, $\pi \in \Pi^M$, *a priori*. In this case, the policy is composed of actions based only on the current state at each decision epoch. Therefore the policy is a distribution over the actions: $\pi = \{\pi_t(s_t) = (\pi_t(1 \mid s_t), \ldots, \pi_t(|\mathcal{A}| \mid s_t)) \in \mathscr{M}(\mathcal{A}) : a_t \in \mathcal{A}, s_t \in \mathcal{S}, t \in \mathcal{T}\}$. In this policy, $\pi_t(a_t \mid s_t)$ is the probability of selecting action $a_t \in \mathcal{A}$ if the MMDP is in state $s_t \in \mathcal{S}$ at time $t \in \mathcal{T}$. Then, after the DM has specified the policy, nature randomly selects model $m \in \mathcal{M}$ with probability λ_m. Now, nature selects $s_1 \in \mathcal{S}$ according to the initial distribution $\mu_1^m \in \mathscr{M}(\mathcal{S})$, and the DM selects an action, $a_1 \in \mathcal{A}$, according to the pre-specified distribution $\pi_1(s_1) \in \mathscr{M}(\mathcal{A})$. Then, nature selects the next state $s_2 \in \mathcal{S}$ according to $p_1^m(\cdot|s_1, a_1) \in \mathscr{M}(\mathcal{S})$. The interaction carries on in this way where the DM selects actions according to the pre-specified policy, π, and nature selects the next state according to the distribution given by the corresponding row of the transition probability matrix. From this point of view, it is easy to see that under a fixed policy, the dynamics of the stochastic process follow a Markov chain. Policy evaluation then is straightforward; one can use backward recursion. While policy evaluation is similar for MMDPs as compared to standard MDPs, policy optimization is much more challenging for MMDPs. For example, value iteration, a well-known solution technique for MDPs, does not apply to MMDPs where actions are coupled across models.

2.3.2. The adaptive problem

The adaptive problem generalizes the non-adaptive problem to allow the DM to utilize realizations of the states to adjust her strategy. In this problem, nature and the DM interact sequentially where the DM gets new information in each decision epoch of the MMDP and the DM is allowed to utilize the realizations of the states to infer information about the ambiguous problem parameters when selecting her future actions. In this setting, nature begins the interaction by selecting a model, $m \in \mathcal{M}$, according to the distribution Λ, and the model selected is not known to the DM. Nature then selects an initial state $s_1 \in \mathcal{S}$ according to the model's initial distribution, μ_1^m. Next, the DM observes the state, s_1, and makes her move by selecting an action, $a_1 \in \mathcal{A}$. At this point, nature randomly samples the next state, $s_2 \in \mathcal{S}$, according to the distribution given by $p_1^m(\cdot|s_1, a_1) \in \mathscr{M}(\mathcal{S})$. The interaction continues by alternating between the DM (who observes the state and selects an action) and nature (who selects the next state according to the distribution defined by the corresponding row of the transition probability matrix).

In the adaptive problem, the DM considers the current state of the MMDP along with information about all previous states observed and actions taken. Because the history is available to the DM, the DM may be able to infer which model is most likely to correctly characterize the behavior of nature which the DM is observing. As we will formally prove later, in this context the DM will specify a history-dependent policy in general, $\pi = \{\pi_t(h_t) : h_t \in \mathcal{S} \times \mathcal{A} \times \ldots \times \mathcal{A} \times \mathcal{S}, t \in \mathcal{T}\}$.

2.4. Analysis of MMDPs

In this section, we will analyze the WVP as defined in (2.4). For both the adaptive and non-adaptive problems, we will describe the classes of policies that achieve the optimal weighted value, the complexity of solving the problem, and related problems that may provide insights into promising solution methods. These results and solution methods are summarized in Table 2.1.

2.4.1. General properties of the weighted value problem

In both the adaptive and non-adaptive problems, nature is confined to the same set of rules. However, the set of strategies available to the DM in the non-adaptive problem is

Property	Non-adaptive Problem		Adaptive Problem	
Always an optimal Markov policy?	Yes	Proposition 2.5	No	Corollary 2.1
Always an optimal deterministic policy?	Yes	Proposition 2.5	Yes	Corollary 2.2
Computational Complexity	NP-hard	Proposition 2.6	PSPACE-hard	Proposition 2.3
Exact Solution Method	MIP	Proposition 2.7	Outer linearization with state-wise pruning	Procedure 2 Procedure 3
Approximation Algorithm	WSU Mean Value Problem	Procedure 1 –	–	–

Table 2.1: Summary of the main properties and solution methods related to the non-adaptive and adaptive problems for MMDPs. Solution methods with dashed entries are not discussed in this thesis.

just a subset of the strategies available in the adaptive problem. Therefore, if W_N^* and W_A^* are the best expected values that the DM can achieve in the non-adaptive and adaptive problems, respectively, then it follows that $W_N^* \leq W_A^*$.

Proposition 2.1. *$W_N^* \leq W_A^*$. Moreover, the inequality may be strict.*

Proof. Consider the MMDP illustrated in Figure 2.2.

First, we describe the decision epochs, states, rewards, and actions for this MMDP. This MMDP is defined for 3 decision epochs where state 1 is the only possible state for decision epoch 1, states 2 and 3 are the states for decision epoch 2, and state 4 is the only state reachable in decision epoch 3. States 5 and 6 are terminal states. This MMDP has two models $\mathcal{M} = \{1, 2\}$. For each model, the only non-zero reward is received upon reaching the terminal state 5. In states 1, 2, and 3, the DM only has one choice of action $a = 1$. In state 4, the DM can select between action $a = 1$ and $a = 2$.

Now we will describe the transition probabilities for each model. Each line represents a transition that happens with probability one when the corresponding action is selected. Solid lines correspond to transitions for model $m = 1$ and dashed lines correspond to transitions for model $m = 2$.

Since state 4 is the only state in which there is a choice of action, we define the possible policies selecting an action in this state. Consider the adaptive problem for this MMDP. The optimal decision rule for state 4 will depend on the state observed at time $t = 2$: If the history of the MMDP is $(s_1 = 1, a_1 = 1, s_2 = 2, a_2 = 1)$, then select action 1, otherwise select action 2. In model 1, the only way to reach state 4 is through state 2. Upon observing this sample path, the policy prescribes taking action 1 which will lead to a transition to state 5 and thus a reward of 1 will be received. On the other hand, in model 2, the only

way to reach state 4 is through state 3. Therefore, the policy will always prescribe taking action 2 in model 2 which leads to state 5 with probability 1. This means that evaluating this policy in model 1 gives an expected value of 1 and evaluating this policy in model 2 gives an expected value of 1. Therefore, for any given weights λ, this policy has a weighted value of $W_A^* = 1$.

Now, consider the non-adaptive problem for the MMDP. Before the DM can observe the state at time $t = 2$, she must specify a decision rule to be taken in state 4. For state 4, there are two options: select action 1 or select action 2. Let q be the probability of selecting action 1. If action 1 is selected, this will give an expected value of 1 in model 1 and an expected value of 0 in model 2, which produces a weighted value of λ_1. Analogously, if action 2 is selected, the weighted value in the MMDP will be λ_2. Thus, the optimal policy for the non-adaptive problem gives a weighted value of $\max_{q\in[0,1]}\{q\lambda_1, (1-q)\lambda_2\}$ which will be exactly $\max\{\lambda_1, \lambda_2\}$.

This means that for any choice of λ such that $\lambda_1 < 1$ and $\lambda_2 < 1$, the MMDP has $W_N^* = \max\{\lambda_1, \lambda_2\} < 1 = W_A^*$. In this MMDP, there does not exist a Markov policy that is optimal for the adaptive problem. □

Corollary 2.1. *It is possible that there are no optimal policies that are Markovian for the adaptive problem.*

The results of Proposition 2.1 and Corollary 2.1 mean that the DM may benefit from being able to recall the history of the MMDP. This history allows for the DM to infer which model is most likely, conditional on the observed sample path and tailor the future actions to reflect this changing belief about nature's choice of model. Therefore, the DM must search for policies within the history-dependent policy class to find an optimal solution to the adaptive MMDP. These results establish that the adaptive problem does not reduce to the non-adaptive problem in general. For this reason, we separate the analysis for the adaptive and non-adaptive problems.

2.4.2. Analysis of the adaptive problem

We begin by establishing an important connection between the adaptive problem and the POMDP [67]:

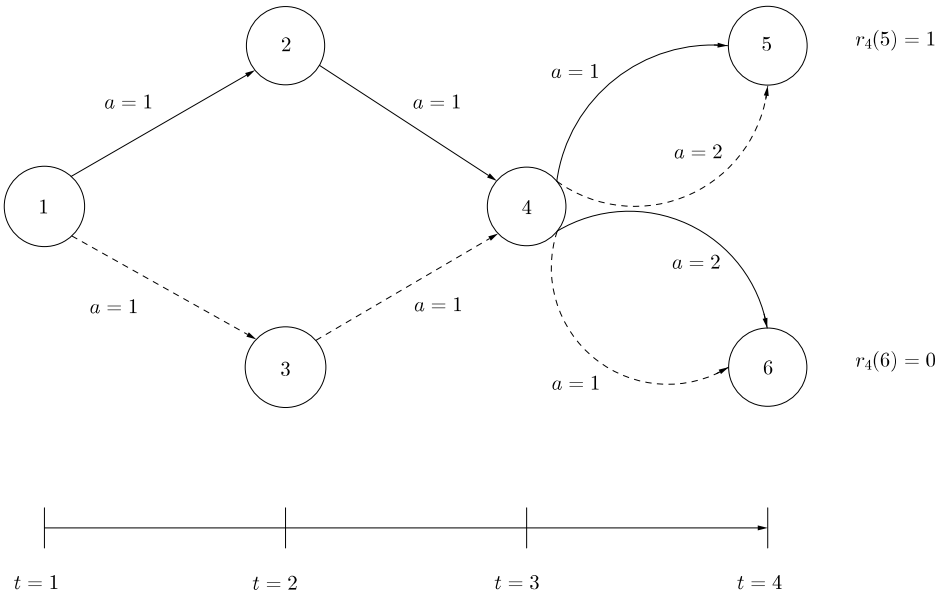


Figure 2.2: An example of an MMDP with $W_A > W_N$. The MMDP shown has six states, two actions, and two models. Each arrow represents a transition that occurs with probability 1 for the corresponding action labeling the arrow. Solid lines represent transitions in model 1 and dashed lines represent transitions in model 2. There are no intermediate rewards in this MMDP, but there is a terminal reward of 1 if state 5 is reached.

Proposition 2.2. *Any MMDP can be recast as a special case of a POMDP such that the maximum weighted value of the MMDP is equivalent to the expected discounted rewards of the POMDP.*

Proof. Let $(\mathcal{T}, \mathcal{S}, \mathcal{A}, \mathcal{M}, \Lambda)$ be an MMDP. From this MMDP, we can construct a POMDP in the following way. The *core states* of the POMDP will be constructed as *state-model* pairs, $(s, m) \in \mathcal{S} \times \mathcal{M}$. The action space for the POMDP is the same as the action space for the MMDP, $\mathcal{A}$. We construct the rewards for the POMDP, denoted r^P, as follows:

$$r^P((s, m), a) := \lambda_m r^m(s, a), \forall s \in \mathcal{S}, m \in \mathcal{M}, a \in \mathcal{A}.$$

The transition probabilities among the core states are defined as follows:

$$p((s', m') \mid (s, m), a) = \begin{cases} p^m(s' \mid s, a) & \text{if } m' = m, \\ 0 & \text{otherwise.} \end{cases}$$

This observation space of the POMDP has a one-to-one correspondence to the state space of the MMDP. We will label the observation space for the POMDP as $\mathcal{O} := \{1, \ldots, S\}$ where $S := |\mathcal{S}|$. In this POMDP, the observations give perfect information about the state element of the state-model pair, but no information about the model element of the state-model pair, and the conditional probabilities are defined accordingly:

$$q(s|(s_t, m)) = \begin{cases} 1 & \text{if } s = s_t, \\ 0 & \text{otherwise.} \end{cases}$$

This special structure on the observation matrix ensures that the same policy is evaluated in each model of the MMDP. By the construction of the POMDP, any history-dependent policy that acts on the sequence of states (observations in the case of the POMDP) and actions $(s_1, a_1, s_2, \ldots, a_{t-1}, s_t)$ will have the same expected discounted rewards value in the POMDP as the weighted value for the MMDP. □

Remark 2.1. *If the state-model pairs that make up the POMDP core state space are ordered as* $(1,1), \ldots, (S,1),\ (1,2), \ldots,\ (S,2), \ldots, (1,M), \ldots, (S,M)$*, then the transition*

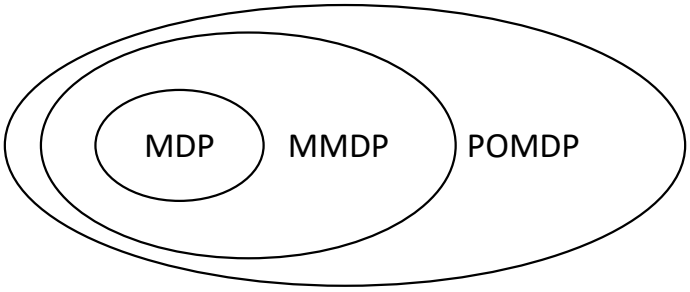


Figure 2.3: A Venn diagram illustrating the relationship between an MDP, MMDP, and POMDP. As shown in Proposition 2.2, any MMDP is a special case of a POMDP due to the structure of the transition matrix and observation conditional probabilities. Further, an MDP is a special case of an MMDP in which the MMDP only has one model.

probability matrix has the following block diagonal structure:

$$P_t(a_t) := \begin{bmatrix} P_t^1(a_t) & 0 & \dots & 0 \\ 0 & P_t^2(a_t) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_t^M(a_t) \end{bmatrix}.$$

The block diagonal structure of the transition probability matrix implies that the underlying Markov chain defined on the core states is reducible.

Corollary 2.2. *There is always a deterministic policy that is optimal for the adaptive problem.*

The implication of Proposition 2.2 is illustrated in Figure 2.3 which displays the relationship between MDPs, MMDPs, and POMDPs. Given Proposition 2.2, we can draw on similar ideas proposed in the literature for solving POMDPs and refine them to take advantage of structural properties specific to MMDPs. This important connection was first established by Le Tallec [40]. However, we show that even though MMDPs have special structure on the observation matrix and transition probability matrix (see the proof of Proposition 2.2), we cannot expect any improvements in the complexity of the problem due to this structure. The following result was first proved by Le Tallec [40] and we subse-

quently did the same analysis independently. We include the proof using our notation for completeness.

Proposition 2.3. *The adaptive problem for MMDPs is PSPACE-hard.*

Proof. This result follows from the original proof of complexity for POMDPs from [54]. Although the MMDP is a special case of a POMDP, we illustrate that the special structure in the observation matrix and transition probabilities is precisely the special case of POMDPs used in the original complexity proof. To aid the reader's understanding, we reproduce the proof here with the modifications to make it specific to MMDPs. We also provide Figure 2.4 which illustrates the construction of an MMDP from the quantified satisfiability problem with two clauses for two existential variables and a universal variable.

First, we assume that $\lambda_m \in (0,1)\ \forall m \in \mathcal{M}$. To show that the adaptive WVP for MMDPs is PSPACE-hard, we reduce QSAT to this problem. We start from any quantified boolean formula $(Q_1 u_1)(Q_2 u_2)\cdots(Q_n u_n)F(u_1, u_2, \ldots, u_n)$ with n variables, n quantifiers (i.e, Q_i is $\exists$ or $\forall$), and m clauses $C_1, C_2, \ldots, C_m$. We construct an MMDP with m models such that its optimal policy has weighted value of 0 or less if and only if the formula is true. The MMDP is constructed as follows: for every variable u_i, we will generate states corresponding to two decision epochs $2i-1$ and $2i$. In decision epoch $2i-1$, there will be two states, A'_i and A_i. In decision epoch $2i$, there will be four states, T'_i, F'_i, T_i, and F_i. After the last decision epoch (at time $2n+1$), there will be 2 states, A_{n+1} and A'_{n+1}. The initial state is A'_1 for every model. The action space is constructed as follows: for every existential variable u_i, the states A'_i and A_i each have two possible actions, $true$ (T) and $false$ (F), which are elements of the action set $\{T, F\}$. All other states have only one action. The models of the MMDP correspond to the clauses in the quantified formula. Each model's transition probabilities are defined as follows: for every existential variable, the transitions out of A'_i and A_i are deterministic according to the action taken. For state A'_i (A_i), selecting action $true$ will ensure that the next state is T'_i (T_i) and selecting action $false$ will ensure that the next state is F'_i (F_i). For every universal variable u_i, the transitions from A'_i (A_i) to T'_i (T_i) and from A'_i (A_i) to F'_i (F_i) occur with equal probability. The differences between the models' transition probabilities occur depending on the negation of variables within the corresponding clause. For every variable u_i that is not negated in the clause, transitions occur deterministically from T'_i to A_{i+1}, F'_i to A'_{i+1}, T_i to A_{i+1}, and F_i to A'_{i+1}. For every variable u_i that is negated in the clause, transitions occur deterministically from T'_i to A'_{i+1}, F'_i to A_{i+1}, T_i to A'_{i+1}, and F_i to A_{i+1}. The initial

state is A'_1 for every model. There is a terminal cost of 1 upon reaching state A'_{n+1} and no cost for reaching A_{n+1}. Other than the terminal costs, there are no costs associated with any of the states or actions.

Now that we have constructed the MMDP, we must show that there exists a policy that achieves a weighted value of zero if and only if the statement is true. First, we show that if there exists a history-dependent policy with a weighted value of zero, then the statement must be true. Consider that such a policy exists. Recall that for every model, we start in state A'_1. In order to achieve a weighted value equal to zero, the policy must ensure that we end in state A_{n+1} for every model. If not, we incur a cost of 1 at time $2n+1$ in one of the models $m \in \mathcal{M}$ which has weight $\lambda_m > 0$, and thus the weighted value is not zero. If we were able to reach state A_{n+1} in every model, this would imply that our policy is able to select actions for states A'_i and A_i for existential variables u_i based on observation of the previous universal variables in a way that the clause is satisfied. Since this occurs for all models, each clause must be true.

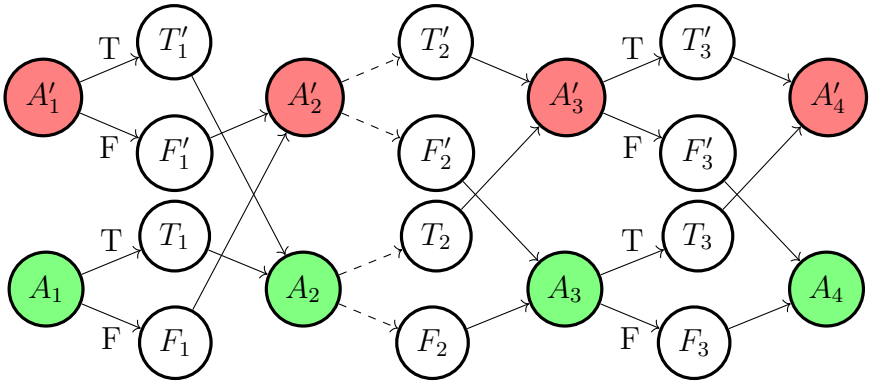
Next, we show that if the quantified formula is true, then there exists a policy that achieves a weighted value of zero. If the quantified formula is true, this means that there exist choices of the existential variables that satisfy the statement. For every existential variable u_i, one can select the appropriate action in $\{T, F\}$ so that based on the values of the previous universal variables, the statement is still true. This corresponds to a policy that will end up in state A_{n+1} with probability one for all models. Thus, this policy achieves a weighted value equal to zero.

□

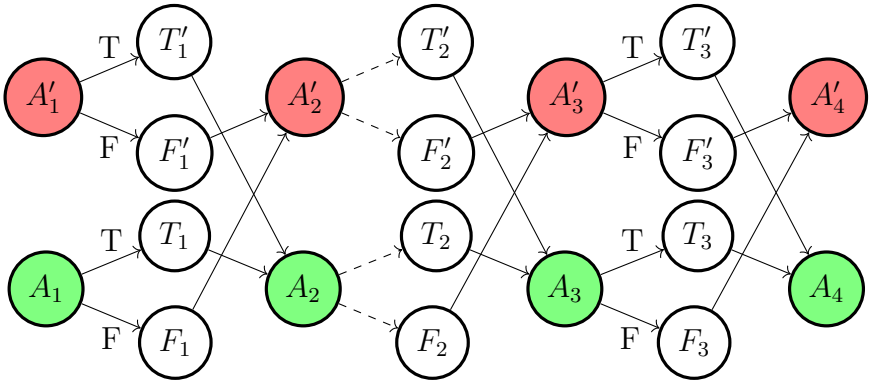
Although the adaptive problem is PSPACE-hard, and we cannot expect to develop an algorithm whose solution time is bounded above by a function that is polynomial in the problem size, we now discuss some special properties of the problem that can be exploited to develop an exact algorithm for solving this problem in Section 2.5. We start by establishing a sufficient statistic for MMDPs:

Definition 2.3 (Information state for MMDPs)**.** *The information state for an MMDP is given by a vector:*

$$b_t := \Big[b_t(1,1), \ldots, b_t(S,1),\ b_t(1,2), \ldots,\ b_t(S,2), \ldots, b_t(1,M), \ldots, b_t(S,M)\Big]'$$



(a) The transitions probabilities in Model 1 represents the first clause over the quantified variables, $u_1 \vee\, !u_2 \vee\, !u_3$.



(b) The transitions probabilities in Model 2 that represents the second clause over the quantified variables, $u_1 \vee u_2 \vee\ u_3$.

Figure 2.4: An example of reduction of QSAT to an adaptive MMDP. The figure illustrates how the quantified formula $\exists u_1 \forall u_2 \exists u_3 (u_1 \vee\, !u_2 \vee u_3) \wedge (u_1 \vee\ u_2 \vee\, !u_3)$ can be represented as an MMDP. Solid lines represent transitions that occur with probability. Dashed lines represent transitions that occur out of the state with equal probability. Transitions corresponding to the actions $true$ and $false$ are labeled with T and F, respectively. State A'_i represents the case where the clause is false at this point and states A_i represents the case where the clause is true at this point.

with elements:

$$b_t(s_t, m) := \mathbb{P}\left(s_t, m \mid s_1, a_1, \ldots, s_{t-1}, a_t, s_t\right).$$

The fact that the information state is a sufficient statistic follows directly from Proposition 2.2, the formulation of a POMDP, and the special structure in the observation matrix.

Given this sufficient statistic, we establish some structural properties of the WVP:

Proposition 2.4. *The information state, b_t, has the following properties:*

1. *The value function is piece-wise linear and convex in the information state, b_t.*
2. *$b_t(s, m) > 0 \Rightarrow b_t(s', m) = 0,\ \forall s' \neq s$.*
3. *The information state as defined above is Markovian in that the information state b_{t+1} depends only on the information state and action at time t, b_t and a_t respectively, and the state observed at time t+1, s_{t+1}.*

Proof of 2.4.1. We will prove this by induction. At time $T+1$, the value function is represented as:

$$v_{T+1}(b_{T+1}) = b'_{T+1} r_{T+1}, \forall b_{T+1} \in B$$

which is linear (and therefore piecewise linear and convex) in b_{T+1}. Now, we perform the induction step. The inductive hypothesis is that the value function at $t+1$ is piecewise linear and convex in b_{t+1} and therefore can be represented by set of hyperplanes $\mathcal{B}$ such that $v_{t+1}(b_{t+1}) = \max_{\beta_{t+1} \in \mathcal{B}_{t+1}} \beta'_{t+1} b_{t+1}$.

$$\begin{aligned}
v_t(b_t) &= \max_{a_t \in \mathcal{A}} \left\{ b'_t r_t(a_t) + \alpha \sum_{s_{t+1} \in \mathcal{S}} \gamma(s_{t+1} | b_t, a_t) v_{t+1}(T(b_t, a_t, s_{t+1})) \right\} \\
&= \max_{a_T \in \mathcal{A}} \left\{ b'_t r_t(a_t) + \alpha \left[\sum_{s_{t+1} \in \mathcal{S}} \left(\sum_{m' \in \mathcal{M}} \sum_{s_t \in \mathcal{S}} p^{m'}(s_{t+1} | s_t, a_t) b_t(s_t, m') \right) \cdot v_{t+1}(T(b_t, a_t, s_{t+1})) \right] \right\} \\
&= \max_{a_t \in \mathcal{A}} \left\{ \sum_{s_t \in \mathcal{S}} \sum_{m \in \mathcal{M}} r_t^m(s_t, a_t) \cdot b_t(s_t, m) \right. \\
&\qquad \left. + \alpha \sum_{s_{t+1} \in \mathcal{S}} \sum_{m \in \mathcal{M}} \max_{\beta_{t+1} \in B_{t+1}} \beta_{t+1}(s_{t+1}, m) \cdot \sum_{s_t \in \mathcal{S}} p^m(s_{t+1} | s_t, a_t) b_t(s_t, m) \right\}
\end{aligned}$$

$$= \max_{a_t \in \mathcal{A}} \left\{ \sum_{s_t \in \mathcal{S}} \sum_{m \in \mathcal{M}} \left(r_t^m(s_t, a_t) + \sum_{s_{t+1} \in \mathcal{S}} \max_{\beta_{t+1} \in \mathcal{B}_{t+1}} \beta_{t+1}(s_{t+1}, m) \cdot p^m(s_{t+1}|s_t, a_t) \right) b_t(s_t, m) \right\} \tag{2.5}$$

which is piece-wise linear and convex in b_t. Therefore, we can represent (2.5) as the maximum over a set of hyperplanes:

$$v_t(b_t) = \max_{\beta_t \in B_t} \{\beta_t' b_t\},$$

where

$$B_t := \{\beta_t : \beta_t = r_t(a) + \alpha P_t'(a)\beta_{t+1}, a \in \mathcal{A}, \beta_{t+1} \in B_{t+1}\}.$$

□

Proof of 2.4.2. This follows directly from the definition of the information state 2.3 and the definition of the conditional probabilities in (2.4.2). To elaborate, we prove this by induction: In the initial decision epoch, s_1 is observed and so for every $m \in \mathcal{M}$, only the state corresponding to (s_1, m) can have a positive value. Now, suppose that at time t, only $|\mathcal{M}|$ values of b_t are positive, and they correspond to the state-model pairs (s, m) with $s = s_t$. Then, the DM selects an action a_t and a new state, s_{t+1}, is observed. At this point, only states (s, m) with $s = s_{t+1}$ can have positive values. □

Proof of 2.4.3. Next, we show that the information state can be efficiently transformed in each decision epoch using Bayesian updating. That is, we aim to show that the information state is Markovian in that the information state at the next stage only depends on the information state in the current stage, the action taken, and the state observed in the next stage.

$$b_{t+1} = T(b_t, a_t, s_{t+1}) \tag{2.6}$$

Consider the information state at time 1 at which point state s_1 has been observed. This information state can be represented by the vector with components:

$$b_1(s, m) = \begin{cases} \dfrac{\lambda_m \mu_1^m(s)}{\sum_{m' \in \mathcal{M}} \lambda_m \mu_1^{m'}(s)} & \text{if } s = s_1 \\ 0 & \text{otherwise} \end{cases}$$

Now, suppose that the information state at time t is b_t, the decision-maker takes action $a_t \in \mathcal{A}$, and observes state s_{t+1} at time $t+1$. Then, every component of the information state can be updated by:

$$b_{t+1}(s,m) = \begin{cases} T^m(b_t, a_t, s_{t+1}) & \text{if } s = s_{t+1} \\ 0 & \text{otherwise} \end{cases}$$

where

$$T^m(b_t, a_t, s_{t+1}) := \frac{\sum_{s_t \in \mathcal{S}} p_t^m(s_{t+1}|s_t, a_t) b_t(s_t, m)}{\sum_{m' \in \mathcal{M}} \sum_{s_t \in \mathcal{S}} p_t^{m'}(s_{t+1}|s_t, a_t) b_t(s_t, m')}$$

which follows from the following:

$$\begin{aligned}
b_{t+1}(s_{t+1}, m) &= \mathbb{P}\left(m \mid h_{t+1}\right) \\
&= \mathbb{P}\left(m \mid s_{t+1}, a_t, h_t\right) && (2.7) \\
&= \frac{\mathbb{P}\left(m, s_{t+1} \mid a_t, h_t\right)}{\mathbb{P}\left(s_{t+1} \mid a_t, h_t\right)} && (2.8) \\
&= \frac{\mathbb{P}\left(s_{t+1} \mid m, a_t, h_t\right) \mathbb{P}\left(m \mid a_t, h_t\right)}{\sum_{m' \in \mathcal{M}} \mathbb{P}\left(s_{t+1} \mid m, a_t, h_t\right) \mathbb{P}\left(m \mid a_t, h_t\right)} && (2.9) \\
&= \frac{\mathbb{P}\left(s_{t+1} \mid m, a_t, h_t\right) \mathbb{P}\left(m \mid h_t\right)}{\sum_{m' \in \mathcal{M}} \mathbb{P}\left(s_{t+1} \mid m, a_t, h_t\right) \mathbb{P}\left(m \mid h_t\right)} && (2.10) \\
&= \frac{\sum_{s_t \in \mathcal{S}} p_t^m(s_{t+1} \mid s_t, a_t) \mathbb{1}(s_t) \mathbb{P}\left(m \mid h_t\right)}{\sum_{m' \in \mathcal{M}} \sum_{s_t \in \mathcal{S}} p_t^{m'}(s_{t+1} \mid s_t, a_t) \mathbb{1}(s_t) \mathbb{P}\left(m' \mid h_t\right)} && (2.11) \\
&= \frac{\sum_{s_t \in \mathcal{S}} p_t^m(s_{t+1} \mid s_t, a_t) b_t(s_t, m)}{\sum_{m' \in \mathcal{M}} \sum_{s_t \in \mathcal{S}} p_t^{m'}(s_{t+1} \mid s_t, a_t) b_t(s_t, m')} && (2.12)
\end{aligned}$$

if $s_{t+1} \in \mathcal{S}$ is in fact the state observed at time $t+1$. (2.7) follows from the definition of h_{t+1}, (2.8) and (2.9) follow from the laws of conditional probability and total probability. (2.10) follows because the action is selected independently of the context. (2.11) follows from the definition of $p^m(s_{t+1} \mid s_t, a_t)$ and an indicator which denotes the state at time t, and (2.12) follows from the definition of the information state at time t. We define the operator T such that the element at (s, m) in $T(b_t, a_t, s_{t+1})$ is exactly $T^m(b_t, a_t, s_{t+1})$ if $s = s_{t+1}$ and 0 otherwise.

Therefore, the information state is Markovian in that the information state at time $t+1$ only relies on the information state at time t, the action taken at time t, and the state observed at time $t+1$. □

According to part 1, the optimal value function can be expressed as the maximum value over a set of hyperplanes. This structural result forms the basis of our exact algorithm in Section 2.5.1. Part 2 states that only elements in the vector with the same value for the state portion of the state-model pair (s, m) can be positive simultaneously, which implies that at most $|\mathcal{M}|$ elements of this vector are zero. This result allows us to ignore the parts of this continuous state space that have zero probability of being occupied. Part 3 allows for a sequential update of the belief that a given model is the best representation of the observed states given the DM's actions according to Bayes' rule. Consider the information state at time 1 at which point state s_1 has been observed. This information state can be represented by the vector with components:

$$b_1(s,m) = \begin{cases} \dfrac{\lambda_m \mu_1^m(s)}{\sum_{m' \in \mathcal{M}} \lambda_m \mu_1^{m'}(s)} & \text{if } s = s_1, \\ 0 & \text{otherwise.} \end{cases}$$

Now, suppose that the information state at time t is b_t, the DM takes action $a_t \in \mathcal{A}$, and observes state s_{t+1} at time $t+1$. Then, every component of the information state can be updated by:

$$b_{t+1}(s,m) = \begin{cases} T^m(b_t, a_t, s_{t+1}) & \text{if } s = s_{t+1}, \\ 0 & \text{otherwise,} \end{cases}$$

where $T^m(b_t, a_t, s_{t+1})$ is a Bayesian update function that reflects the probability of model m being the best representation of the system given the most recently observed state, the previous action, and the previous belief state:

$$T^m(b_t, a_t, s_{t+1}) := \frac{\sum_{s_t \in \mathcal{S}} p_t^m(s_{t+1}|s_t, a_t) b_t(s_t, m)}{\sum_{m' \in \mathcal{M}} \sum_{s_t \in \mathcal{S}} p_t^{m'}(s_{t+1}|s_t, a_t) b_t(s_t, m')}.$$

As mentioned previously, our focus in this article is on applications of the MMDP framework to medical problems in contexts for which learning by Bayesian updating is not appropriate. However, the adaptive framework would apply to other contexts. We describe solution methods that exploit these structural properties in Section 2.5.1.

2.4.3. Analysis of the non-adaptive problem

In this section, we analyze the non-adaptive problem for which restricts the DM's policy is restricted to the class of Markov policies (Π^M). We begin by establishing the important result that there always exists a deterministic optimal policy for the special case of the non-adaptive problem. This result is important because searching among policies in the Markov deterministic policy class may be appealing for several reasons: First, each individual model is solved by a policy in this class and it could be desirable to find a policy with the same properties as the each model's individual optimal policy. Second, Markov policies are typically easier to implement because they only require the current state to be stored rather than partial or complete histories of the MDP. Third, Markov deterministic policies are ideal for medical decision making, the motivating application for this article, because they can be easily translated to treatment guidelines that are based solely on the information available to the physician at the time of the patient visit, such as the patient's current blood pressure levels. For applications in medicine, such as the case study in Section 2.7, deterministic policies are a necessity since randomization is unlikely to be considered ethical outside the context of randomized clinical trials.

Proposition 2.5. *For the non-adaptive problem, there is always a Markov deterministic policy that is optimal.*

Proof. Let μ_t^π be the probability distribution induced over the states by the partial policy used up to time t in the MMDP, so that $\mu_t^\pi(s_t, m) = P(s_t \mid \pi_{1:(t-1)})$, where $\pi_{1:(t-1)}$ is the partial policy over decision epochs 1 through $(t-1)$. Now we will prove the proposition by induction on the decision epochs.

The base case of the proof is the last decision epoch, T: For any partial policy $\pi_{1:(T-1)}$, there will be some stochastic process that induces the probability distribution μ_T^π. Given μ_T^π, the best decision rules are found by:

$$\begin{aligned}
&\max_q \sum_{s_T \in \mathcal{S}} \max \sum_{a_T \in \mathcal{A}} q_t(a_T|s_T) \cdot \\
&\qquad \left(\sum_{m \in \mathcal{M}} \mu_T^\pi(s_T, m) \left[r_T^m(s_T, a_T) + \sum_{s_{T+1}} p^m(s_{T+1}|s_T, a_T) r_{T+1}^m(s_{T+1}) \right] \right) \\
&\quad \text{s.t. } q_T(a_T|s_T) \geq 0, \quad \forall s_T \in \mathcal{S}, a_T \in \mathcal{A}, \\
&\qquad \sum_{a_T \in \mathcal{A}} q_T(a_T|s_T) = 1, \forall s_T \in \mathcal{S}.
\end{aligned}$$

Since we are selecting the action probabilities independently for each state, we can focus on the maximization problem:

$$\max_{q_T(s_T)} \sum_{a_T \in \mathcal{A}} q_t(a_T|s_T) \sum_{m \in \mathcal{M}} \mu_T^{\pi}(s_T, m) \left[r_T^m(s_T, a_T) + \sum_{s_{T+1}} p^m(s_{T+1}|s_T, a_T) r_{T+1}^m(s_{T+1}) \right]$$

$$\text{s.t. } q_T(a_T|s_T) \geq 0,$$

$$\sum_{a_T \in \mathcal{A}} q_T(a_T|s_T) = 1,$$

which is a LP, and will have a solution where at most 1 action has a non-zero value of $q_T(a_T|s_T)$. Thus, for any given partial policy $\pi = (\pi_1, \ldots, \pi_{T-1})$, the optimal decision rule at time T will be deterministic.

Next, we assume that for any partial policy $\pi_{1:t} = (\pi_1, \pi_2, \ldots, \pi_t)$, there exists deterministic decision rules that are optimal for the remainder of the horizon: $\pi^*_{(t+1):T} = (\pi^*_{t+1}, \pi^*_{t+2}, \ldots, \pi^*_T)$, and that the partial beginning policy used up to decision epoch t, $(\pi_1, \ldots, \pi_{t-1})$, has induced the probability distribution μ_t^{π}. We will show that it follows that there exists a deterministic decision rule that is optimal for decision epoch t:

$$\sum_{s_t \in \mathcal{S}} \max_{q} \sum_{a_t \in \mathcal{A}} q_t(a_t|s_t) \sum_{m \in \mathcal{M}} \mu_t^{\pi}(s_t, m) \left[r_t^m(s_t, a_t) + \sum_{s_{t+1}} p^m(s_{t+1}|s_t, a_t) v_{t+1}^m(s_{t+1}) \right]$$

$$\text{s.t. } q_t(a_t|s_t) \geq 0,$$

$$\sum_{a_t \in \mathcal{A}} q_t(a_t|s_t) = 1.$$

Once again, we can focus on the maximization problem within the sum:

$$\max_{q_t(a_t|s_t)=1} \sum_{a_t \in \mathcal{A}} q_t(a_t|s_t) \sum_{m \in \mathcal{M}} \mu_t^{\pi}(s_t, m) \left[r_t^m(s_t, a_t) + \sum_{s_{t+1}} p^m(s_{t+1}|s_t, a_t) v_{t+1}^m(s_{t+1}) \right]$$

$$\text{s.t. } q_t(a_t|s_t) \geq 0,$$

$$\sum_{a_t \in \mathcal{A}} q_t(a_t|s_t) = 1.$$

This is a LP so there will exist an extreme point solution that is optimal. This extreme point solution corresponds to a deterministic decision rule for decision epoch t. □

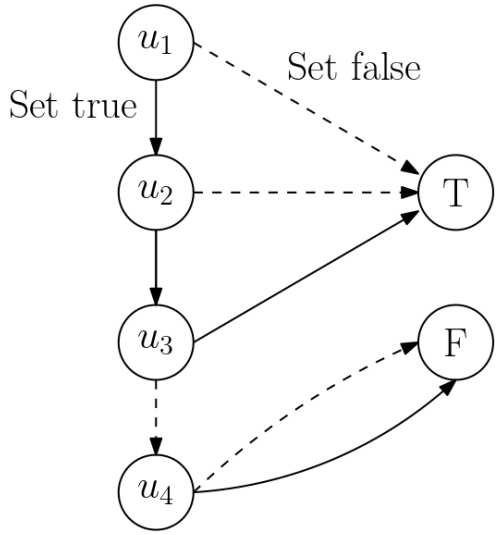
This result means that for the non-adaptive problem, the DM can restrict her attention to the class of Markov deterministic policies. This result may be surprising at first due to the result of Fact 2 in [65] which states that the best stationary randomized policy can be arbitrarily better than the best stationary deterministic policy for POMDPs. While this result may seem to contradict Proposition 2.5, it is worth noting that Fact 2 of [65] was derived in the context of an infinite-horizon MDP in which it is possible that the same state can be visited more than once. In the finite-horizon MMDP, it is not possible that s_t could be visited more than once.

Even though the non-adaptive problem requires searching over a smaller policy class than for the adaptive problem ($\Pi^{MD} \subset \Pi^{HD}$), the non-adaptive problem is still provably hard. The following result was first proved by Le Tallec [40] under the name of MDP with random uncertainty. We subsequently did the same analysis independently and include the proof using our notation for completeness.

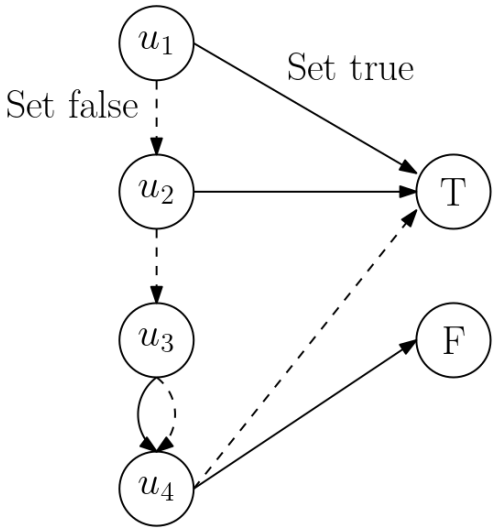
Proposition 2.6. *Solving the non-adaptive problem for an MMDP is NP-hard.*

Proof. We show that any 3-CNF-SAT problem can be transformed into the problem of determining if there exists a Markov deterministic policy for an MMDP such that the weighted value is greater than zero. Let's suppose we have a 3-CNF-SAT instance: a set of variables $U = \{u_1, u_2, \ldots, u_n\}$ and a formula $E = C_1 \wedge C_2 \ldots \wedge C_m$. We will construct an MMDP with one decision epoch from this instance of 3-CNF-SAT. In the only decision epoch, the state space consists of one state per variable, $u_i,\ i = 1, \ldots, n$. At the terminal stage, there are two states labeled "T" and "F". There are no immediate rewards for this problem. For every state u_i, there are two actions $true$ or $false$. The terminal rewards correspond to a cost of 0 for reaching the terminal state "T" and a cost of 1 upon reaching the terminal state "F".

The transition probabilities for model j correspond to the structure of clause C_j and are defined as follows: for any variable $u_i,\ i < n$ that does not appear in Clause j, both actions lead to the state u_{i+1} with probability 1. If variable u_n does not appear in Clause j, both actions lead to the state "F" with probability 1. For any variable u_i that appears non-negated in clause C_j, the action $true$ leads from state u_i to state "T" with probability 1 and the action $false$ leads from state u_i to state u_{i+1} with probability 1. For the variables that appear negated in the clause, the action $true$ leads from state u_i to state u_{i+1} with probability 1 and the action $false$ leads from state u_i to state "T" with probability 1. The initial distribution of all models is variable u_1 with probability 1.



(a) The transitions probabilities in model 1 that represent the first clause: $C_1 = !u_1 \vee !u_2 \vee u_3$.



(b) The transitions probabilities in model 2 that represent the second clause: $C_2 = u_1 \vee u_2 \vee !u_4$.

Figure 2.5: An example of the reduction of 3-CNF-SAT to a non-adaptive MMDP. The figures illustrates how a 3-CNF-SAT instance, $E = (u_1 \vee !u_2 \vee u_3) \wedge (u_1 \vee u_2 \vee !u_4)$, can be represented as an MMDP. Solid lines represent the transitions associated with the action *true* and dashed lines represent the transitions associated with the action *false*. All transitions shown happen with probability 1.

We will show that there is a truth assignment for the variables in U that satisfies E if and only if there is a Markov deterministic policy for the MMDP that achieves a weighted value equal to 0.

First, we show that if there is a truth assignment for the variables in U that satisfies E, then there exists a Markov deterministic policy for the MMDP that achieves a weighted value equal to 0. To construct such a policy, take the action $true$ in every state u_i such that u_i is true is the satisfying truth assignment and take the action $false$ otherwise. Because this true assignment satisfies each clause, the corresponding policy will reach state "T" with probability 1 in each model. By construction, this policy will have a weighted value of zero.

Next, we show that if there is a policy $\Pi = \Pi^{MD}$ that achieves a weighted value of 0, that there exists a truth assignment that will satisfy E. Suppose that policy $\pi \in \Pi^{MD}$ achieves a cost of zero. This implies that for every clause, the policy π leads to the state "T" with probability 1. We can construct a truth assignment from this policy by assigning u_i to be true if $\pi(u_i)$ is $true$, and u_i to be false if $\pi(u_i)$ is $false$.

Therefore, we have created a one-to-one mapping of truth assignments to MD policies such that any policy that satisfies E will also have weighted value 0. Hence, if we were able to find a policy that achieves a weighted value of 0 in polynomial time, we would also be able to solve 3-CNF-SAT in polynomial time. Thus, the MMDP WVP with $\Pi = \Pi^{MD}$ is NP-hard. □

The result of Proposition 2.6 implies that we cannot expect to find an algorithm that solves the non-adaptive problem for all MMDPs in polynomial time. Still, we are able to solve the non-adaptive problem by formulating it as an MIP as discussed in the following proposition.

Proposition 2.7. *Non-adaptive MMDPs can be formulated as the following MIP:*

$$\max_{\pi,\, v} \quad \sum_{m\in\mathcal{M}} \sum_{s\in\mathcal{S}} \lambda_m \mu_1^m(s) v_1^m(s) \tag{2.13a}$$

$$\text{s.t.} \quad \sum_{a\in\mathcal{A}} \pi_t(a|s) = 1, \qquad \forall s \in \mathcal{S}, t \in \mathcal{T}, \tag{2.13b}$$

$$M\pi_t(a|s) + v_t^m(s) - \sum_{s'\in\mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \le r_t^m(s,a) + M, \tag{2.13c}$$

$$\forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T},$$

$$v^m_{T+1}(s) \leq r^m_{T+1}(s), \forall m \in \mathcal{M}, s \in \mathcal{S}, \tag{2.13d}$$

$$\pi_t(a|s) \in \{0,1\}, \quad \forall a \in \mathcal{A}, s \in \mathcal{S}, t \in \mathcal{T}. \tag{2.13e}$$

Proof. The decision variable $v_t(s)$ represents the optimal value-to-go for state $s \in \mathcal{S}$ at time $t \in \mathcal{T}$. The dual variables correspond to the probability of selecting an action given a state. Corner point solutions correspond to deterministic policies, and the optimal policy is deterministic by construction.

For an MMDP, we cannot use the standard LP formulation used to solve MDPs because of the requirement that the policy must be the same in each of the different models. The mixed-integer program shown in (2.13) gives a formulation that ensures that the policy $\pi \in \Pi^{MD}$ is the same in each model. Each decision variable, $v^m_t(s)$ represents the value-to-go from state $s \in \mathcal{S}$ at time $t \in \mathcal{T}$ for model $m \in \mathcal{M}$ corresponding to the policy $\pi \in \Pi^{MD}$ that maximizes the weighted value of the MMDP. To enforce that the same policy in each model, $m \in \mathcal{M}$, we introduce binary decision variables, $x_{s,a,t}$ for every state, $s \in \mathcal{S}$, action, $a \in \mathcal{A}$, and decision epoch $t \in \{1, 2, \ldots, T\}$. If $x_{s,a,t}$ takes on a value of 1, this means that the best policy dictates taking action a in state s at time t for every model, and $x_{s,a,t} = 0$ otherwise. If the choice of M is sufficiently large (e.g., $M > (|\mathcal{T}| + 1) \cdot \max_{m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T}} r_t(s,a)$), then the inequalities will become tight when the corresponding binary decision variable $x_{s,a,t} = 1$, because all of the other actions' constraints will have a large value, M, added to their value in the second inequality. The equality constraint ensures that every state-time pair only has one action prescribed. □

In this formulation, the decision variables, $v^m_t(s) \in \mathbb{R}$, represent the value-to-go from state $s \in \mathcal{S}$ at time $t \in \mathcal{T}$ in model $m \in \mathcal{M}$. The binary decision variables, $\pi_t(a|s) \in \{0,1\}$, take on a value of 1 if the policy prescribes taking action $a \in \mathcal{A}$, in state $s \in \mathcal{S}$, at epoch $t \in \mathcal{T}$, and 0 otherwise.

It is well-known that standard MDPs can be solved using a LP formulation [56, §6.9]. Suppose that $v_t(s,a)$ represents the value-to-go from state $s \in \mathcal{S}$ using action $a \in \mathcal{A}$ at decision epoch $t \in \mathcal{T}$. The LP approach for solving MDPs utilizes a reformulation trick that finding $\max_{a \in \mathcal{A}} v_t(s,a)$ is equivalent to finding $\min v_t(s)$ such that $v_t(s) \geq v_t(s,a)$ for all feasible a. In this reformulation, the constraint $v_t(s) \geq v_t(s,a)$ is tight for all actions that are optimal. The MIP formulation presented in (2.13) relies on similar ideas as the LP formulation of an MDP, but is modified to enforce the constraint that the policy must be the same across all models.

In the MIP formulation of the non-adaptive MMDP, we require that constraints

$$v_t^m(s) \leq r_t^m(s,a) + \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') + M(1 - \pi_t(a|s)), \quad \forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}$$

are tight for the action $a^* \in \mathcal{A}$ such that $\pi_t(a^*|s) = 1$ for any given state $s \in \mathcal{S}$, decision epoch $t \in \mathcal{T}$, and model $m \in \mathcal{M}$. The purpose of the big-M is to ensure that $v_t^m(s) = v_t^m(s,a)$ only if $\pi_t(a|s) = 1$ meaning that the value-to-go for this state-time pair in model $m \in \mathcal{M}$ corresponds to the policy that is being used in all models. Thus, if action $a \in \mathcal{A}$ is selected (and thus, $\pi_t(a|s) = 1$), we want $v_t^m(s) = v_t^m(s,a)$ and if not ($\pi_t(a|s) = 0$), we want $v_t^m(s) \leq v_t^m(s,a)$. Therefore, we must select M sufficiently large enough for all constraints. We also consider formulations of the MIP based on the dual LP formulation of an MDP (see Puterman [56]) in Appendix A. The formulation (A.1) is suited for Lagrangian methods such as progressive hedging, and the formulation (A.6) is suited to nonlinear programming methods. These additional formulations could be useful for future research on alternative methods for solving MMDPs.

The formulation of the non-adaptive problem as an MIP may seem more natural after a discussion of the connections with two-stage stochastic programming [9]. If we view the non-adaptive problem through the lens of stochastic programming, the $\pi_t(a|s)$ binary variables that define the policy can be interpreted as the *first-stage decisions* of a two-stage stochastic program. Moreover, nature's choices of model, $\mathcal{M}$, correspond to the possible *scenarios* which are observed according to the probability distribution Λ. In this interpretation, the value function variables, $v_t^m(s)$, can be viewed as the *recourse decisions*. That is, once the DM has specified the policy according to the π variables and nature has specified a model $m \in \mathcal{M}$, the DM seeks to maximize the value function so long as it is consistent with the first-stage decisions:

$$V(\pi, m) = \max_{\pi} \Big[v^m(\pi) \mid \sum_{a \in \mathcal{A}} \pi_t(a|s) = 1, \\ \forall s \in \mathcal{S}, t \in \mathcal{T}, \pi_t(a|s) \in \{0,1\}, \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T} \Big],$$

where $V(\pi)$ is the *recourse function*. This can be written as

$$V(\pi) = \mathbb{E}^m \left[V(\pi, m) \right] = \mathbb{E}^{\pi, P^m, \mu_1^m} \left[\sum_{t=1}^{T} r_t(s_t, a_t) + r_{T+1}(s_{T+1}) \right].$$

The formulation in (2.13) is the deterministic equivalent formulation of this stochastic integer program.

Our initial numerical experiments showed that moderate-sized MDPs can be solved using (2.13), but this approach may be too computationally intensive to solve large problems such as those that arise in the context of medical decision making. This motivated the development of an approximation algorithm that we describe in Section 2.5, subsequently test on randomly generated problem instances in Section 2.6, and then apply to a medical decision making problem in the case study in Section 2.7. The following relaxation of the non-adaptive problem allows us to quantify the performance of our approximation algorithm:

Proposition 2.8. *For any policy $\hat{\pi} \in \Pi$, the weighted value is bounded above by the weighted sum of the optimal values in each model. That is,*

$$\sum_{m \in \mathcal{M}} \lambda_m v^m(\hat{\pi}) \leq \sum_{m \in \mathcal{M}} \lambda_m \max_{\pi \in \Pi^{MD}} v^m(\pi), \quad \forall \hat{\pi} \in \Pi.$$

Proof. The result follows from this series of inequalities:

$$\begin{aligned} \sum_{m \in \mathcal{M}} \lambda_m v^m(\hat{\pi}) &\leq \max_{\pi \in \Pi^{MD}} \sum_{m \in \mathcal{M}} \lambda_m v^m(\pi) \qquad (2.14) \\ &\leq \sum_{m \in \mathcal{M}} \lambda_m \max_{\pi \in \Pi^{MD}} v^m(\pi), \end{aligned}$$

where (2.14) states than any MD policy will have a weighted value at most the optimal MD policy's weighted value. This optimal weighted value, in turn, is at most the value that can be achieved by solving each model separately and then weighting these values. □

The result of Proposition 2.8 allows us to evaluate the performance of any MD policy even when we cannot solve the WVP exactly to determine the true optimal policy. We use this result to illustrate the performance of our approximation algorithm in Section 2.7.

Proposition 2.8 motivates several connections between robustness and the value of information. First, the upper bound in Proposition 2.8 is based on the well-known *wait-and-see* problem in stochastic programming that relaxes the condition that all models must have the same policy. Second, the expected value of perfect information (EVPI) is the expected

value of the wait-and-see solution minus the recourse problem solution:

$$EVPI = \left[\sum_{m\in\mathcal{M}} \lambda_m \max_{\pi\in\Pi^M} v^m(\pi)\right] - \max_{\pi\in\Pi^M}\left[\sum_{m\in\mathcal{M}} \lambda_m v^m(\pi)\right].$$

While the wait-and-see value provides an upper bound, it may prescribe a set of solutions, one for each model, and thus it often does not provide an implementable course of action. Another common approach in stochastic programming is to solve the MVP which is a simpler problem in which all parameters take on their expected values. In the MMDP, this corresponds to the case where all transition probabilities and rewards are weighted as follows:

$$\bar{p}_t(s'|s,a) = \sum_{m\in\mathcal{M}} \lambda_m p_t^m(s'|s,a), \quad \forall s\in\mathcal{S}, a\in\mathcal{A}, t\in\mathcal{T}$$

and

$$\bar{r}_t(s,a) = \sum_{m\in\mathcal{M}} \lambda_m r_t^m(s,a).$$

Solving the MVP will give a single policy, $\bar{\pi}$, which we will term the *mean value solution*, with the following expected rewards:

$$W(\bar{\pi}) = \sum_{m\in\mathcal{M}} \lambda_m v^m(\bar{\pi}).$$

Thus, we can measure the robustness in an MMDP via the value of the stochastic solution (VSS):

$$VSS = W^* - W(\bar{\pi}),$$

which is a common measure of the impact of randomness in stochastic programming [9, §4.2]. If VSS is low, this implies that there is not much value from solving the MMDP versus the MVP. On the other hand, if VSS is high, this implies that the DM will benefit significantly from solving the MMDP.

While the non-adaptive problem has connections to stochastic programming, it also has connections to POMDPs. The non-adaptive problem can be viewed as the problem of finding the best *memoryless controller* for this POMDP [72]. Memoryless controllers for POMDPs are defined on the most recent observation only. For an MMDP, this would translate to the DM specifying a policy that is based only on the most recent observation of the state (recall that the DM gets no information about the model part of the state-model

pair). Because no history is allowed to be incorporated into the definition of the policy, this policy is permissible for the non-adaptive problem. These connections between MMDPs and stochastic programs and POMDPs allow us to better understand the complexity and potential solution methods for finding the best solution to the non-adaptive problem.

2.5. Solution methods

2.5.1. Solution methods for the adaptive problem

In this section, we present an exact solution method that can be used to solve the adaptive problem for an MMDP. We begin by describing Procedure 1 which is an exact solution method for solving the adaptive weighted value problem. The correctness of this solution method follows from Proposition 2.2 which states that every MMDP is a special case of a POMDP and that the maximum weighted value is equivalent to the expected discounted rewards of the corresponding POMDP. Therefore, we transform the MMDP into a POMDP and use a solution method analogous to a well-known solution method for POMDPs [67]. This method exploits the property that the value function is piece-wise linear convex and therefore can be represented as the maximum over a set of supporting hyperplanes (Proposition 2.4).

In the worst-case, the number of hyperplanes needed to represent the value function could potentially be as large as $1 + |\mathcal{A}| + \sum_{t=1}^{T-1} |\mathcal{A}|^{|\mathcal{S}|+T-t}$ for $T \geq 2$, but in many cases the number of hyperplanes that are actually needed to represent the optimal value function is much smaller. *Pruning* describes the methods by which hyperplanes that are not needed to represent the optimal value function are discarded. The pruning method described in Procedure 2 is based on the LP method described in [67], but exploits the result of Proposition 2 for computational gain. This result states that only certain parts of the information space are reachable due to the special structure of the MMDP, and this allows for the LP problems for pruning to be decomposed into a set of smaller LPs.

For this procedure, we will use the information state as defined in Definition 2.3 and define the following notation:

$$r_{T+1}^m := \begin{bmatrix} r_{T+1}(1) \\ \vdots \\ r_{T+1}(|\mathcal{S}|) \end{bmatrix}, r_t^m(a_t) := \begin{bmatrix} r_t(1, a_t) \\ \vdots \\ r_t(|\mathcal{S}|, a_t) \end{bmatrix}, \forall m \in \mathcal{M}, \forall a_t \in \mathcal{A},$$

$$r_{T+1} := \begin{bmatrix} r_{T+1}^1 \\ \vdots \\ r_{T+1}^{|\mathcal{M}|} \end{bmatrix}, \qquad r_t(a_t) := \begin{bmatrix} r_t^1(a_t) \\ \vdots \\ r_t^{|\mathcal{M}|}(a_t) \end{bmatrix}, \forall a_t \in \mathcal{A},$$

For every action, we define the block diagonal matrix:

$$P_t(a_t) := \begin{bmatrix} P_t^1(a_t) & 0 & \dots & 0 \\ 0 & P_t^2(a_t) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_t^M(a_t) \end{bmatrix},$$

where each matrix $P_t^m(a_t)$, $\forall m \in \mathcal{M}$ is the transition probability matrix in decision epoch $t \in \mathcal{T}$ associated with action $a_t \in \mathcal{A}$ for model $m \in \mathcal{M}$. The matrix Q represents the analog of the conditional probability matrix for observations:

$$Q := [\underbrace{I_{|\mathcal{S}|}, \dots, I_{|\mathcal{S}|}}_{|\mathcal{M}| \text{ times}}]',$$

where $I_{|\mathcal{S}|}$ denotes an $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix. We use $Q(s_t)$ to denote the column vector corresponding to $s_t \in \mathcal{S}$ such that the elements indexed (s, m) in this vector have values

$$q(s_t|(s,m)) = \begin{cases} 1 & \text{if } s = s_t \\ 0 & \text{otherwise} \end{cases}$$

for all $m \in \mathcal{M}$.

The space of all information states at time t is

$$B_t = \left\{ b_t \ : \ b_t(s,m) \geq 0, \ \forall (s,m) \in \mathcal{S} \times \mathcal{M}, \ \sum_{m \in \mathcal{M}} b_t(s,m) = 1, \forall s \in \mathcal{S} \right\}.$$

Procedure 1 is a backwards induction algorithm which generates a set of hyperplanes at each decision epoch. Procedure 2 eliminates hyperplanes that are not necessary to represent the optimal value function. The DM selects the optimal sequence of actions for the observed history in an analogous way to a POMDP: update the information state based on the observation and select the action corresponding to the maximizing hyperplane at this particular information state.

Algorithm 1: Algorithm for solving the WVP (2.2) for an adaptive MMDP]

Data: MMDP
Result: Collection $B_0, \ldots, B_T$
Initialize $\mathcal{B}_{T+1} = \{r_{T+1}\}$
The value-to-go at time $T+1$. $v_{T+1}(b_{T+1}) = \beta'_{T+1} b_{T+1}$, $\forall b_{T+1} \in B_{T+1}$
$t \leftarrow T$
while $t \geq 0$ **do**
  for *Every action* a_t **do**

$$\mathcal{B}_t(a_t) \leftarrow \Big\{ \beta_t(a_t) : \beta_t(a_t) = r_t(a_t) + \sum_{s_{t+1} \in \mathcal{S}} P_t(a_t) \mathbf{diag}(Q(s_{t+1})) \beta_{t+1}^{s_{t+1}}, \quad \forall \beta_{t+1}^1 \times \cdots \times \beta_{t+1}^{|\mathcal{S}|} \in \mathcal{B}_{t+1} \times \cdots \times \mathcal{B}_{t+1} \Big\}$$

  end
  $\mathcal{B}_t \leftarrow \cup_{a_t \in \mathcal{A}} \mathcal{B}_t(a_t)$
  State-wise Prune$(\mathcal{B}_t)$
  The value-to-go at time t is $v_t(b_t) = \max_{\beta_t \in \mathcal{B}_t} \beta'_t b_t, \quad \forall b_t \in B_t$
  $t \leftarrow t - 1$
end

We do not consider approximation algorithms for the adaptive problem in this thesis. There is an extensive literature on approximation algorithms for POMDPs [42], and the results we presented for exact solution methods could also be adapted for use in some of those approximation algorithms.

2.5.2. Solution methods for the non-adaptive problem

In this section, we will discuss how to leverage the results of Section 2.4 to solve the non-adaptive problem. We discuss the MIP formulation of Proposition 2.7 for solving the non-adaptive weighted value problem. Although the MIP formulation provides a viable way to exactly solve this class of problems, the result of Proposition 2.6 motivates the need for a fast approximation algorithm that can scale to large MMDPs

Algorithm 2: State-wise Prune

Data: A set of vectors in $\mathbb{R}^{|\mathcal{S}\times\mathcal{M}|}$, $\mathcal{B}$.
Result: $\mathcal{B}$

for *Every vector* $\beta \in \mathcal{B}$ **do**
  for *Every state* $s \in \mathcal{S}$ **do**
    Let $\mathcal{B}(s) = \{\beta_s : \beta_s(m) = \beta(s,m), \beta \in \mathcal{B}\}$
    Solve the LP (2.15)

$$
\begin{aligned}
z_s^* := \min_{\mu_s \in \mathscr{M}(\mathcal{M}), x \in \mathbb{R}} \quad & x - \beta_s' \mu_s && (2.15)\\
\text{s.t.} \quad & x \geq \bar{\beta}_s' \mu_s \qquad \forall \bar{\beta}_s \in \mathcal{B}(s),\\
& \sum_{m \in \mathcal{M}} \mu_s(m) = 1
\end{aligned}
$$

    If $\prod_{s\in\mathcal{S}} z_s^* > 0$, remove β from $\mathcal{B}$
  end
end

Mixed-integer programming formulation

The big-M constraints are an important aspect of the MIP formulation of the weighted value problem. Thus, we discuss tightening of the big-M values in the following constraints:

$$
\begin{aligned}
v_t^m(s) \leq r_t^m(s,a) + \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') + M(1 - \pi_t(a|s)),\\
\forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \{1, \ldots, T\}.
\end{aligned}
$$

Recall that the decision variables of the form $v_t^m(s) \in \mathbb{R}$ represent the value-to-go from state $s \in \mathcal{S}$ at time $t \in \mathcal{T}$ in model $m \in \mathcal{M}$ under the policy specified by the π variables. For the purposes of this discussion, we define the optimal value function for epoch t and model m for a given state-action pair (s,a) as:

$$
\begin{aligned}
v_t^m(s,a) = r_t^m(s,a) + \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') + M(1 - \pi_t(a|s)),\\
\forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \{1, \ldots, T\}.
\end{aligned}
$$

For action $a \in \mathcal{A}$, we would like the smallest value of M's that still ensure that:

$$r_t^m(s,a) + \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \leq r_t^m(s,a') + \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a') v_{t+1}^m(s') + M_{m,s,t}, \ \forall a' \in \mathcal{A}.$$

Rearranging, we obtain:

$$M_{m,s,t} \geq r_t^m(s,a) + \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') - r_t^m(s,a') - \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a') v_{t+1}^m(s'), \quad \forall a, a' \in \mathcal{A}. \tag{2.16}$$

A sufficient condition for (2.16) is the following:

$$M_{m,s,t} \geq \max_{a \in \mathcal{A}} v_t^m(s,a) - \min_{a \in \mathcal{A}} v_t^m(s,a).$$

By the definition of $v_t(s,a)$, we are assuming that the policy defined by the x variables is being followed after time t. However, we can relax this assumption further and allow each model to follow a different policy to obtain the big-M values, where $\max_{a \in \mathcal{A}} v_t^m(s,a)$ is the largest value-to-go for this model and $\min_{a \in \mathcal{A}} v_t^m(s,a)$ is the smallest value-to-go for this model. This will provide tighter bounds that strengthen the MIP formulation, and furthermore these bounds can be computed efficiently using standard dynamic programming methods.

Weight-Select-Update (WSU) Approximation Algorithm

Next, we discuss our Weight-Select-Update (WSU) algorithm, formalized in Procedure 3, which is a fast approximation algorithm for the non-adaptive problem. WSU generates decision rules $\hat{\pi}_t \in \Pi_t^{MD}$ stage-wise starting at epoch T and iterating backwards. At epoch $t \in \mathcal{T}$, the algorithm has an estimate of the value for this policy in each model conditioned on the state s_{t+1} at epoch $t+1 \in \mathcal{T}$. This estimate is denoted $\hat{v}_{t+1}^m(s_{t+1}), \ \forall m \in \mathcal{M}, \forall s_{t+1} \in \mathcal{S}$. The algorithm weights the immediate rewards plus the value-to-go for each of the models and then the algorithm selects, for each state, an action that maximizes the sum of these weighted terms and denotes this action $\hat{\pi}_t(s_t)$. Next, the algorithm updates the estimated value-to-go for every state in each model according to the decision rule $\hat{\pi}_t$ at epoch $t \in \mathcal{T}$. This procedure iterates backwards stage-wise until the actions are specified for the first decision epoch.

Upon first inspection, it may not be obvious that WSU is not guaranteed to produce

Algorithm 3: Weight-Select-Update (WSU) approximation algorithm for the non-adaptive problem (2.4)

Data: MMDP
Result: The policy $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_T) \in \Pi^{MD}$

Let $\hat{v}^m_{T+1}(s_{T+1}) = r^m_{T+1}(s_{T+1}), \ \ \forall m \in \mathcal{M}$
$t \leftarrow T$
while $t \geq 1$ **do**
  for *Every state* $s_t \in \mathcal{S}$ **do**

$$\hat{\pi}_t(s_t) \leftarrow \underset{a_t \in \mathcal{A}}{arg\,max} \Big\{ \sum_{m \in \mathcal{M}} \lambda_m \Big(r^m_t(s_t, a_t) + \sum_{s_{t+1} \in \mathcal{S}} p^m_t(s_{t+1}|s_t, a_t) \hat{v}^m_{t+1}(s_{t+1}) \Big) \Big\} \tag{2.17}$$

  end
  for *Every model* $m \in \mathcal{M}$ **do**

$$\hat{v}^m_t(s_t) \leftarrow r^m_t(s_t, \hat{\pi}_t(s_t)) + \sum_{s_{t+1} \in \mathcal{S}} p^m_t(s_{t+1}|s_t, \hat{\pi}_t(s_t)) \hat{v}^m_{t+1}(s_{t+1}) \tag{2.18}$$

  end
  $t \leftarrow t - 1$
end

the optimal MD policy; however, this approximation algorithm fails to account for the fact that, under a given policy, the likelihood of occupying a specific state could vary under the different models. The result of Proposition 2.9 shows that ignoring this could lead to sub-optimal selection of actions as illustrated in the proof.

Proposition 2.9. *WSU is not guaranteed to produce an optimal solution to the non-adaptive weighted value problem.*

Proof. Consider the counter-example illustrated in Figure 2.6 for $\lambda_1 = 0.8, \lambda_2 = 0.2$. The MMDP has 5 states, 2 actions, 2 models, and 2 decision epochs. First, we can explicitly enumerate all possible deterministic policies for the non-adaptive weighted value problem.

By explicitly enumerating all of the possible deterministic policies, we see that selecting action 1 for state A and action 1 for state B leads to the maximum expected weighted value of $0.9\lambda_2 = 0.72$. Now, consider the resulting policy generated from WSU. There is only one option for state C, so WSU will select $\pi(C) = 1$ and update the value for each

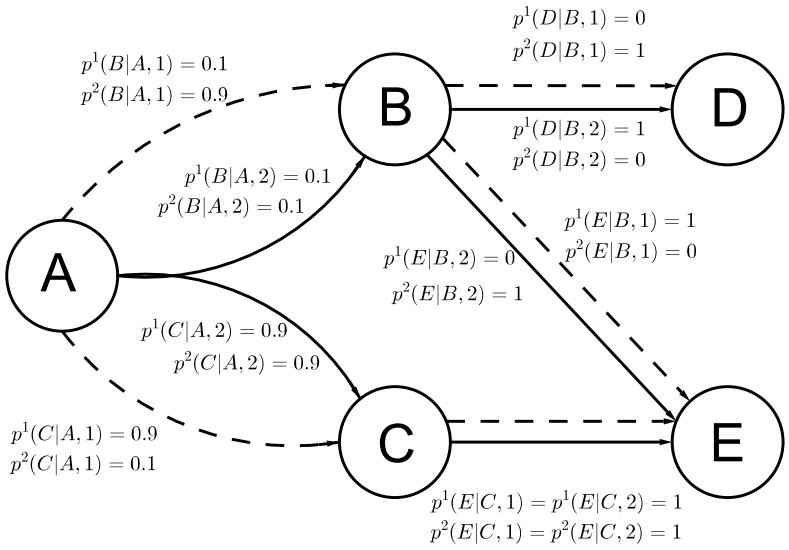


Figure 2.6: An illustration of an MMDP for which the WSU approximation algorithm does not generate an optimal solution to the non-adaptive weighted value problem. Possible transitions for actions 1 and 2 are illustrated with the dashed and solid line respectively. The probability of each possible transition in both of the models is listed by the corresponding line. The DM receives a reward of 1 if state D is reached. Otherwise, no rewards are received.

Table 2.2: An explicit enumeration of the weighted value under every possible deterministic policy for the non-adaptive weighted value problem.

Policy		Expected Values		
State A	**State B**	**Value in Model 1**	**Value in Model 2**	**Weighted Value**
1	1	0	0.9	$0.9\lambda_2 = 0.72$
1	2	0.1	0	$0.1\lambda_1 = 0.08$
2	1	0	0.1	$0.1\lambda_2 = 0.02$
2	2	0.1	0	$0.1\lambda_1 = 0.08$

model as $v^1(C) = 0$ and $v^2(C) = 0$. For state B, WSU will select:

$$\hat{\pi}(B) \leftarrow \underset{a\in\{1,2\}}{arg\,max}\{\lambda_1 p^1(D|B,a) + \lambda_2 p^2(D|B,a)\}$$

and because $\lambda_1 > \lambda_2$, the algorithm will select $\pi(B) = 2$, and then update $v^1(B) = 1$ and $v^2(B) = 0$. Then, the algorithm will select an action for state A as:

$$\hat{\pi}(B) \leftarrow \underset{a\in\{1,2\}}{arg\,max}\{\lambda_1 p^1(B|A,a)\};$$

and so, the algorithm is indifferent between action 1 and action 2 because both give $\lambda_1 p^1(B|A,a) = 0.1\lambda_1$. Therefore, the policy resulting from WSU is either $\hat{\pi} = \{\hat{\pi}(A) = 1, \hat{\pi}(B) = 2, \hat{\pi}(C) = 1\}$ or $\hat{\pi} = \{\hat{\pi}(A) = 2, \hat{\pi}(B) = 2, \hat{\pi}(C) = 1\}$, both of which give a weighted value of $0.1\lambda_1$ which is suboptimal. This shows that WSU may generate a policy that is suboptimal for the non-adaptive weighted value problem. □

Although WSU is not guaranteed to select the optimal action for a given state-time pair, this procedure is guaranteed to correctly evaluate the value-to-go in each model for the procedure's policy, $\hat{\pi}$. This is because, although the action selection in equation (2.17) may be suboptimal, the update of the value-to-go in each model in (2.18) correctly evaluates the performance of this action in each model conditional on being in state s_t at decision epoch t. That is, for a fixed policy, policy evaluation for standard MDPs applies to each of the models, separately.

Proposition 2.10. *For $|\mathcal{M}| = 2$, if $\lambda_m^1 > \lambda_m^2$, then the corresponding policies $\hat{\pi}(\lambda^1)$ and $\hat{\pi}(\lambda^2)$ generated via WSU for these values will be such that*

$$v^m(\hat{\pi}(\lambda^1)) \geq v^m(\hat{\pi}(\lambda^2)).$$

Proof. For ease of notation, we refer to $\hat{\pi}(\lambda^1)$ as π^1. The value-to-go under policy π in model m from state s will be denoted as $v_t^m(s,\pi)$. Because $|\mathcal{M}| = 2$, we will refer to the two models as m and $\bar{m}$ where λ_m is the weight on model m and $(1-\lambda_m)$ is the weight on model $\bar{m}$.

Suppose the proposition is not true; that is, suppose there exists $\lambda_m^1 > \lambda_m^2$ such that $v^m(\hat{\pi}(\lambda^1)) < v^m(\hat{\pi}(\lambda^2))$. Then, it must be the case that for some $t \in \mathcal{T},\ s \in \mathcal{S}$ that

$$v_t^m(s,\pi^1) < v_t^m(s,\pi^2). \tag{2.19}$$

Let t be the last decision epoch in which $\pi_t^1(s_t) \neq \pi_t^2(s_t)$. Note that this implies that $v_{t'}^m(s',\pi^1)) = v_{t'}^m(s',\pi^2), \quad \forall t' > t, s' \in \mathcal{S}$.

First, consider the weighted value problem for $\lambda_m = \lambda_m^1$. Consider a state s at time t for which $\pi_t^1(s) \neq \pi_t^2(s)$. Because the approximation algorithm selected $\pi_t^1(s)$ as the action, it must be that:

$$\begin{aligned} &\lambda_m^1 v_t^m(s,\pi^1) + (1-\lambda_m^1) v_t^{\bar{m}}(s,\pi^1) \geq \lambda_m^1 v_t^m(s,a) + (1-\lambda_m^1) v_t^{\bar{m}}(s,a), \quad \forall a \in \mathcal{A} \\ \Rightarrow &\lambda_m^1 v_t^m(s,\pi^1) + (1-\lambda_m^1) v_t^{\bar{m}}(s,\pi^1) \geq \lambda_m^1 v_t^m(s,\pi^2) + (1-\lambda_m^1) v_t^{\bar{m}}(s,\pi^2) \end{aligned} \tag{2.20}$$

Next, consider the weighted value problem for $\lambda_m = \lambda_m^2$. In this case, for the same state s as above, it must be that the approximation algorithm selected action $\pi_t^2(s)$ because:

$$\begin{aligned} &\lambda_m^2 v_t^m(s,\pi^2) + (1-\lambda_m^2) v_t^{\bar{m}}(s,\pi^2) \geq \lambda_m^2 v_t^m(s,a) + (1-\lambda_m^2) v_t^{\bar{m}}(s,a), \quad \forall a \in \mathcal{A} \\ \Rightarrow &\lambda_m^2 v_t^m(s,\pi^2) + (1-\lambda_m^2) v_t^{\bar{m}}(s,\pi^2) \geq \lambda_m^2 v_t^m(s,\pi^1) + (1-\lambda_m^2) v_t^{\bar{m}}(s,\pi^1). \end{aligned} \tag{2.21}$$

Rearranging (2.20), we have

$$\lambda^1 (v_t^m(s,\pi^1) - v_t^m(s,\pi^2)) + (1-\lambda_m^1)(v_t^{\bar{m}}(s,\pi^1) - v_t^{\bar{m}}(s,\pi^2)) \geq 0, \tag{2.22}$$

and rearranging (2.21), we have

$$\lambda_m^2 (v_t^m(s,\pi^2) - v_t^m(s,\pi^1)) + (1-\lambda_m^2)(v_t^{\bar{m}}(s,\pi^2) - v_t^{\bar{m}}(s,\pi^1)) \geq 0 \tag{2.23}$$

$$\Rightarrow -\lambda_m^2 \left(v_t^m(s,\pi^1) - v_t^m(s,\pi^2)\right) - (1-\lambda_m^2) \left(v_t^{\bar{m}}(s,\pi^1) - v_t^{\bar{m}}(s,\pi^2)\right) \geq 0. \tag{2.24}$$

Adding (2.22) and (2.24), we have:

$$\begin{aligned}
&(\lambda_m^1 - \lambda_m^2)\left(v_t^m(s,\pi^1) - v_t^m(s,\pi^2)\right) + \left((1-\lambda_m^1) - (1-\lambda_m^2)\right)\left(v_t^{\bar{m}}(s,\pi^1) - v_t^{\bar{m}}(s,\pi^2)\right) \geq 0 \\
\Rightarrow &(\lambda_m^1 - \lambda_m^2)\left(v_t^m(s,\pi^1) - v_t^m(s,\pi^2) + v_t^{\bar{m}}(s,\pi^2) - v_t^{\bar{m}}(s,\pi^1)\right) \geq 0. \qquad (2.25)
\end{aligned}$$

Because $\lambda_m^1 > \lambda_m^2$, it must be that

$$\begin{aligned}
&v_t^m(s,\pi^1) - v_t^m(s,\pi^2) + v_t^{\bar{m}}(s,\pi^2) - v_t^{\bar{m}}(s,\pi^1) \geq 0 \\
\Rightarrow &v_t^{\bar{m}}(s,\pi^2) - v_t^{\bar{m}}(s,\pi^1) \geq v_t^m(s,\pi^2) - v_t^m(s,\pi^1) \\
\Rightarrow &v_t^{\bar{m}}(s,\pi^2) > v_t^{\bar{m}}(s,\pi^1), \qquad (2.26)
\end{aligned}$$

where (2.26) follows because of (2.19). However, because $v_t^m(s,\pi^1) < v_t^m(s,\pi^2)$ and $v_t^{\bar{m}}(s,\pi^1) < v_t^{\bar{m}}(s,\pi^2)$, this implies that

$$\lambda_m^1 v_t^m(s,\pi^1) + (1-\lambda_m^1) v_t^{\bar{m}}(s,\pi^1) < \lambda^1 v_t^m(s,\pi^2) + (1-\lambda_m^1) v_t^{\bar{m}}(s,\pi^2),$$

which contradicts that the approximation algorithm would have selected action $\pi_t^1(s)$ for the weighted value problem with $\lambda_m = \lambda_m^1$. Therefore, it must be the case that if $\lambda_m^1 > \lambda_m^2$, then

$$v^m(\hat{\pi}(\lambda^1)) \geq v^m(\hat{\pi}(\lambda^2)).$$

□

Proposition 2.10 guarantees that the policies generated using WSU will have values in model $m \in \mathcal{M}$ that are non-decreasing model m's weight, λ_m. This result is desirable because it allows DMs to know that placing more weight on a particular model will not result in a policy that does worse with respect to that model. Proposition 2.10 is also useful for establishing the lower bound in the following proposition:

Proposition 2.11. *For $|\mathcal{M}| = 2$, any policy generated via WSU will be such that*

$$W(\hat{\pi}(\lambda)) \geq \lambda v^1(\pi^2) + (1-\lambda) v^2(\pi^1).$$

where π^m is the optimal policy for model m.

Proof. Let λ be the weight on model 1, π^1 be an optimal policy for model 1, and π^2 be an

optimal policy for model 2. Due to the result of Proposition 2.10, it follows that

$$v^1(\hat{\pi}(\lambda)) \geq v^1(\pi^2) = v^1(\hat{\pi}(0)), \qquad \forall \lambda \in [0,1],$$
$$v^2(\hat{\pi}(\lambda)) \geq v^2(\pi^1) = v^2(\hat{\pi}(1)), \qquad \forall \lambda \in [0,1],$$

and therefore,

$$W(\hat{\pi}(\lambda)) = \lambda v^1(\hat{\pi}(\lambda)) + (1-\lambda)v^2(\hat{\pi}(\lambda)) \geq \lambda v^1(\pi^2) + (1-\lambda)v^2(\pi^1).$$

□

Proposition 2.11 provides a lower bound on the weighted value of the policy obtained via WSU. While Proposition 2.8 provides an upper bound that applies for any policy $\pi \in \Pi$, the lower bound in Proposition 2.11 is specific to the Markov deterministic policies obtained via WSU for the specific case of 2 models. The bound is generated by appropriately weighting the value obtained by model 1's optimal policy in model 2 and the value obtained by model 2's optimal policy in model 1. This establishes an easy way to obtain this bound because it involves solving the 2 models independently and then evaluating these policies.

2.6. Computational experiments

In this section, we describe a set of computational experiments for comparing solution methods for the adaptive problem and the non-adaptive problem on the basis of run-time and quality of the solution. Our experiments were based on a series of random instances of MMDPs. To generate the random test instances, first the number of states, actions, models, and decision epochs for the problem were defined. Then, model parameters were randomly sampled. In all test instances, it was assumed that the sampled rewards were the same across models, the weights were uninformed priors on the models, and the initial distribution was a discrete uniform distribution across the states. The rewards were sampled from the uniform distribution: $r(s,a) \sim U(0,1), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. The transition probabilities were obtained by sampling from a uniform distribution so that $\tilde{p}^m(s'|s,a) \sim U(0,1)$. Then, for every $(m,s,a,s') \in \mathcal{M} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, the transition probabilities were normalized so that the row of the transition probability matrix had

elements that sum to one:

$$p^m(s'|s,a) := \frac{\tilde{p}^m(s'|s,a)}{\sum_{s''\in\mathcal{S}} \tilde{p}^m(s''|s,a)}.$$

To solve the adaptive version of these instances, Procedure 1 with pruning was used. Procedure 1 was implemented using Python using SciPy's `linprog` package to solve the LPs in the pruning procedure described in Procedure 2. Procedure 1 was terminated if $|\mathcal{B}| > 10,000$. The non-adaptive problem was solved using WSU, MVP, and the MIP formulation. WSU and MVP were implemented using Python 3.5.2. All MIPs were solved using AMPL Version 20150815 and CPLEX 12.6.1.

2.6.1. Comparison of adaptive solution and the non-adaptive solution

Our experiments investigated the time required to solve a set of random instances of MMDPs with 2 states, 2 actions, 2 models for 2 to 5 decision epochs. For each choice of decision epochs, 30 random instances were generated for a total of 120 random instances. For each instance, the non-adaptive problem was solved using the MIP formulation, and the adaptive problem was solved using Procedure 1 with pruning. The non-adaptive problems could be solved in less than 1 second on average and could be solved in less than 10 seconds in the worst-case. Even for the small problems we considered, the time required to solve the adaptive problem could be quite large. For example, the time for Procedure 1 with pruning to solve an adaptive problem with 2 states, 2 actions, 2 models, and 5 decision epochs was over 19 minutes in the worst case and 89 seconds on average. These experiments illustrated that exactly solving the adaptive problem is not a scalable approach, which is consistent with the complexity results of Section 2.4. Moreover, these experiments revealed a very small gap between the adaptive and non-adaptive solutions. For these instances, the average gap (calculated as $\frac{W_A^* - W_N^*}{W_A^*} \times 100\%$) was less than 0.1%, and the worst-case gap was less than 2%.

Due to the solution times in the experiments described above, we did not expect to be able to solve larger instances of the adaptive problem in a reasonable amount of time. To investigate the gap between the solutions of the non-adaptive and the adaptive problems for larger problem sizes, we compared the non-adaptive solution obtained via the MIP to the upper bound from Proposition 2.8 to obtained an upper bound on this value. A base case problem size of 4 states, 4 actions, 4 models, and 4 decision epochs was defined. A

variety of problem sizes were tested by changing one aspect of the base case problem size at a time. The number of states was varied from 4 to 10, the actions from 4 to 10, the models from 4 to 10, for a total of 28 different problem sizes. For each problem size, 100 instances were generated for a total of 2800 random instances. Over these 2,800 random instances, the worst-case gap between the MIP solution and the upper bound was 5.01%, and the average gap was 0.46%. This suggests the benefits of solving the adaptive problem over the non-adaptive problem are small at best. Furthermore, the upper bound from Proposition 2.8 can be used to bound the gap between the non-adaptive solution and the adaptive solution.

2.6.2. Comparison of solution methods for the non-adaptive problem

For the non-adaptive problem, we compared the exact and approximate solution methods on the basis of run-time and quality of solution. To do so, a base case problem size of 4 states, 4 actions, 4 models, and 4 decision epochs was defined. Then, the size of the problem was varied with respect to the number of states, actions, models, and decision epochs independently to determine the influence of growth in the problem size on the average- and worst-case run times and optimality gaps.

To evaluate the quality of the solutions obtained via the WSU approximation algorithm and MVP, we will compare the weighted value policies obtained via the approximation algorithms ($W_N(\hat{\pi})$) to the optimal value obtained by solving the MIP to within 1% of optimality, W_N^*:

$$\text{Gap} = \frac{W_N^* - W_N(\hat{\pi})}{W_N^*} \times 100\%,$$

where $\hat{\pi}$ is the policy obtained from either WSU or MVP. For each problem size tested, the WSU approximation algorithm had a worst-case optimality gap of 1.0% and an average optimality gap being less than 0.01%. The performance of MVP had a worst-case optimality gap of 51.9% and an average gap of 3.5%. These results indicate that the WSU approximation algorithm is likely a better approximation method than MVP.

We also compared the time required to solve the instances using the WSU approximation algorithm and the MIP (see Figure 2.7). The WSU approximation algorithm was able to generate an policy relatively quickly on these test instances (under 1 CPU second on average) while the average time required to solve the MIP noticeably increases as the size of the problem increases, especially with respect to the number of decision epochs in the

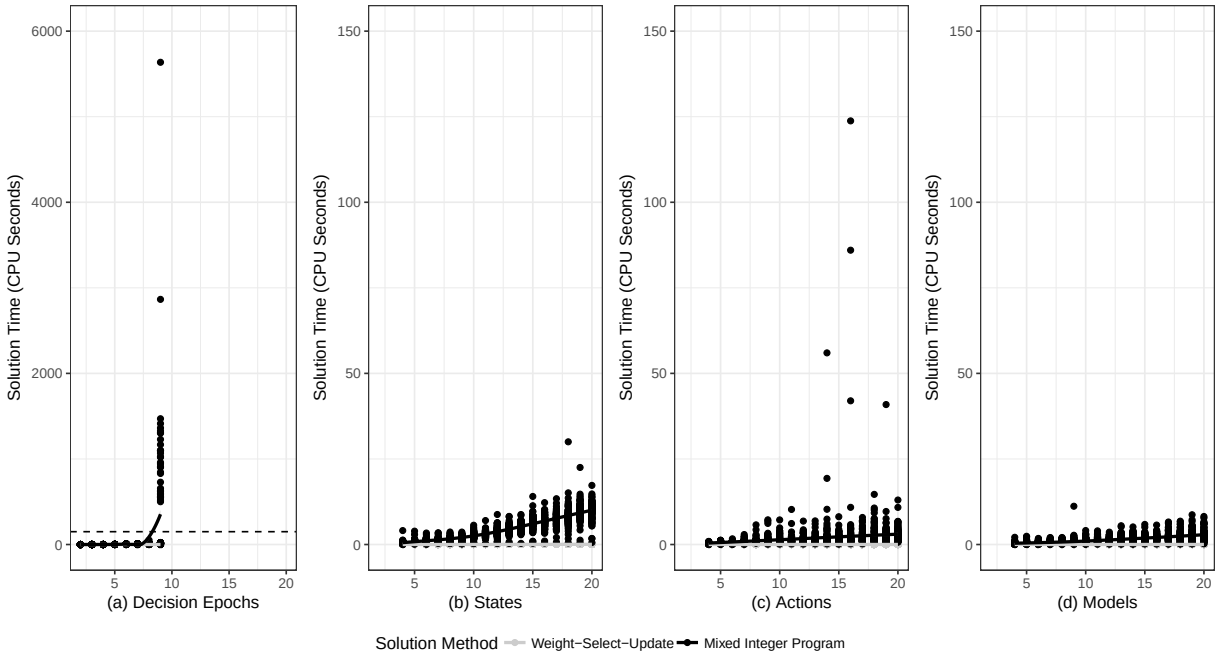


Figure 2.7: The solution times required to solve the non-adaptive problem using the WSU and MIP solution methods for 100 random instances of various problem sizes, in CPU seconds. The base case problem size is 4 decision epochs, 4 states, 4 actions, and 4 models. One aspect of the problem size was varied at a time and 100 random instances were generated and solved using the WSU approximation algorithm and the MIP formulation. The dashed line at 150 CPU seconds in Figure 2.7(a) is used to highlight the change in scale in the y-axis between 2.7(a) and 2.7(b), (c), and (d).

MMDP. These results suggest that approximation algorithm may be needed to approximate solutions for larger MMDPs, such as the one presented as a case study in Section 2.7.

2.7. Case study: blood pressure and cholesterol management for cardiovascular disease prevention in type 2 diabetes

In this section, we present an MMDP to optimize the timing and sequencing of the initiation of blood pressure medications and cholesterol medications for patients with type 2 diabetes. Here, WSU was used to generate a policy that trades off conflicting estimates of cardiovascular risk from two well-established studies in the medical literature. We be-

gin by providing some context about the problem, the MMDP model, and the parameter ambiguity that motivates its use.

Diabetes is one of the most common and costly chronic medical conditions, affecting more than 25 million adults, or 11% of the adult population in the United States [15]. Diabetes is associated with the inability to properly metabolize blood glucose (blood sugar) and other metabolic risk factors that place the patient at risk of complications including coronary heart disease (CHD) and stroke. There are several types of diabetes including type 1 diabetes, in which the patient is dependent on insulin to live, gestational diabetes, which is associated with pregnancy, and type 2 diabetes in which the patient has some ability (albeit impaired) to manage glucose. In this case study we focus on type 2 diabetes, which accounts for more than 90% of all cases.

The first goal, glycemic control, is typically achieved quickly following diagnosis of diabetes using oral medications and/or insulin. Management of cardiovascular risk, the focus of this case study, is a longer term challenge with a complex tradeoff between the harms of medication and the risk of future CHD and stroke events. Patients with diabetes are at much higher risk of stroke and CHD events than the general population. Well-known risk factors include total cholesterol (TC), high density lipoprotein (HDL) often referred to as "good cholesterol", and systolic blood pressure (SBP). Like blood glucose, the risk factors of TC, HDL, and SBP are also controllable with medical treatment. Medications, such as *statins* and *fibrates*, can reduce TC and increase HDL. Similarly, there are a number of medications that can be used to reduce blood pressure including *ACE inhibitors*, *ARBs*, *beta blockers*, *thiazide*, and *calcium channel blockers*. All of these medications have side effects that must be weighed against the long-term benefits of lower risk of CHD and stroke. An added challenge to deciding when and in what sequence to initiate medication is due to the conflicting risk estimates provided by two well known clinical studies: the FHS [75, 76] and the ACC/AHA assessment of cardiovascular risk [27].

2.7.1. MMDP formulation

The MDP formulation of [47] was adapted to create an MMDP based on the FHS risk model [75, 76] and the ACC/AHA risk model [27]. These are the most well-known risk models used by physicians in practice. The state space of the MMDP is a finite set of health states defined by SBP, TC, HDL, and current medications. A discrete set of actions represent the initiation of the two cholesterol medications and 4 classes of blood pressure medications.

The objective is to optimize the timing and sequencing of medication initiation to maximize quality-adjusted life years (QALYs). QALYs are a common measure used to assess health interventions that account for both the length of a patient's life as well as the loss of quality of life due to the burden of medical interventions. For this case study, we will assume that the rewards are the same in each of the models of the MMDP and that only the transition probabilities vary across models. Figure 2.8 provides a simplified example to illustrate the problem. In the diagram, solid lines illustrate the actions of initiating one or both of the most common medications (statins (ST), ACE inhibitors (AI)), and dashed lines represent the occurrence of an adverse event (stroke or CHD event), or death from other causes. In each medication state, including the no medication state ($\emptyset$), patients probabilistically move between health risk states, represented by L (low), M (medium), H (high), and V (very high). For patients on one or both medications, the resulting improvements in risk factors reduce the probability of complications. Treatment *actions* are taken at a discrete set of decision epochs indexed by $t \in \mathcal{T} = \{0, 1, \ldots, T\}$ that correspond to ages 54 through 74 at one year intervals that represent annual preventive care visits with a primary care doctor. States can be separated into *living states* and *absorbing states*. Each living state is defined by the factors that influence a patient's cardiovascular risk: the patient's TC, HDL, and SBP levels, and medication state. We denote the set of the TC states by $\mathcal{L}_{\text{TC}} = \{L, M, H, V\}$, with similar definitions for HDL, $\mathcal{L}_{\text{HDL}} = \{L, M, H, V\}$, and SBP, $\mathcal{L}_{\text{SBP}} = \{L, M, H, V\}$. The thresholds for these ranges are based on established clinically-relevant cut points for treatment [23]. The complete set of health states is indexed by $\ell \in \mathcal{L} = \mathcal{L}_{\text{TC}} \times \mathcal{L}_{\text{HDL}} \times \mathcal{L}_{\text{SBP}}$.

The set of medication states is $\mathcal{M} = \{\tau = (\tau_1, \tau_2, \ldots, \tau_n) : \tau_i \in \{0, 1\}, \forall i = 1, 2, \ldots, 6\}$ corresponding to all combinations of the 6 medications mentioned above. If $\tau_i = 0$, the patient is not on medication i, and if $\tau_i = 1$, the patient is on medication i. The treatment effects for medication i are denoted by $\omega^{\text{TC}}(i)$, for the proportional reduction in TC, $\omega^{\text{HDL}}(i)$, for the proportional change in HDL, and $\omega^{\text{SBP}}(i)$, for the proportional change in SBP, as reported in [47]. The living states in the model are indexed by $(\ell, \tau) \in \mathcal{L} \times \mathcal{M}$. The absorbing states are indexed by $d \in \mathcal{D} = \{\mathcal{D}_{\text{S}}, \mathcal{D}_{\text{CHD}}, \mathcal{D}_{\text{O}}\}$ represent having a stroke, $\mathcal{D}_{\text{S}}$, having a CHD event, $\mathcal{D}_{\text{CHD}}$, or dying, $\mathcal{D}_{\text{O}}$. The action space depends on the history of medications that have been initiated in prior epochs. For each medication, at each epoch,

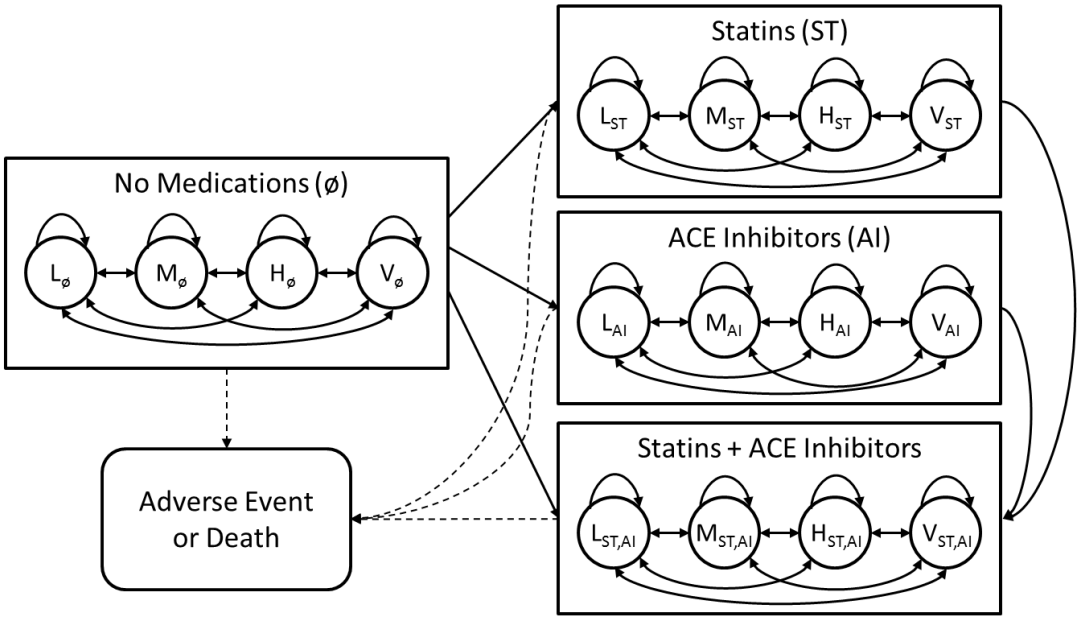


Figure 2.8: An illustration of the state and action spaces of the CVD management MDP (as illustrated in Mason et al. [47]). In the corresponding MMDP, when medications are initiated (solid lines denote actions), the risk factors are improved and the probability of an adverse event (denoted by the dashed lines) is reduced. The probabilities of adverse events may differ in the different models depending on the risk calculator that was used to estimate the probability.

medication i can be initiated (I) or initiation can be delayed (W):

$$A_{(\ell,m_i)} = \begin{cases} \{I_i, W_i\} & \text{if } \tau_i = 0, \\ \{W_i\} & \text{if } \tau_i = 1, \end{cases}$$

and $\mathbf{A}_{(\ell,\tau)} = \{A_{(\ell,\tau_1)} \times A_{(\ell,\tau_2)} \times \cdots \times A_{(\ell,\tau_n)}\}$. Action $\mathbf{a} \in \mathbf{A}_{(\ell,\tau)}$ denotes the action in state (ℓ, τ). If a patient is in living state (ℓ, τ) and takes action $\mathbf{a}$, the new medication state is denoted by τ', where τ_i' is set to 1 for any medications i that are newly initiated by action $\mathbf{a}$; $\tau_i' = \tau_i$ for all medications i which are not newly initiated. Once medication i is initiated, the associated risk factor is modified by the medication effects denoted by $\omega^{\mathrm{TC}}(i)$, $\omega^{\mathrm{HDL}}(i)$, and $\omega^{\mathrm{SBP}}(i)$, resulting in a reduction in the probability of a stroke or CHD event. Two types of transition probabilities are incorporated into the model: probabilities of transition among health states and the probability of events (fatal and nonfatal). At epoch t, $\bar{p}_t^{\tau}(d|\ell)$ denotes the probability of transition from state $(\ell, \tau) \in \mathcal{L} \times \mathcal{M}$ to an absorbing state $d \in \mathcal{D}$. Given that the patient is in health state $\ell \in \mathcal{L}$, the probability of being in health state ℓ' in the next epoch is denoted by $q_t(\ell'|\ell)$. The health state transition probabilities, $q_t(\ell'|\ell)$, were computed from empirical data for the natural progression of BP and cholesterol adjusted for the absence of medication [20]. We define $p_t^{\tau}(j|\ell)$ to be the probability of a patient being in state $j \in \mathcal{L} \cup \mathcal{D}$ at epoch $t+1$, given the patient is in living state (ℓ, τ) at epoch t. The transition probabilities can be written as:

$$p_t^{\tau}(j|i) = \begin{cases} \big[1 - \sum_{d \in \mathcal{D}} \bar{p}_t^{\tau}(d|i)\big] q_t(j|i) & \text{if } i, j \in \mathcal{L}, \\ \bar{p}_t^{\tau}(j|i) & \text{if } i \in \mathcal{L},\ j \in \mathcal{D}, \\ 1 & \text{if } i = j \in \mathcal{D}, \\ 0 & \text{otherwise.} \end{cases}$$

The two models of the MMDP represent the different cardiovascular risk calculators used to estimate the transition probabilities to the absorbing states: $\bar{p}_t^{\tau}(d|i)$ for $i \in \mathcal{L}, d \in \mathcal{D}$. We will refer to the model using the ACC/AHA study as model A and the model using FHS as model F. We weight these models by $\lambda_A \in [0, 1]$ and $\lambda_F := 1 - \lambda_A$ respectively. We estimate of all other cause mortality based take from the Centers for Disease Control and Prevention life tables [3]. The reward $r_t(\ell, \tau)$ for a patient in health state ℓ at epoch

t is:

$$r_t(\ell, \tau) = \mathcal{Q}(\ell, \tau),$$

where $\mathcal{Q}(\ell, \tau) = 1 - d^{\text{MED}}(\tau)$ is the reward for one QALY. QALYs are elicited through patient surveys, and are commonly used for health policy studies [29]. The *disutility* factor, $d^{\text{MED}}(\tau)$, represents the estimated decrease in quality of life due to the side effects associated with the medications in τ.

2.7.2. Results

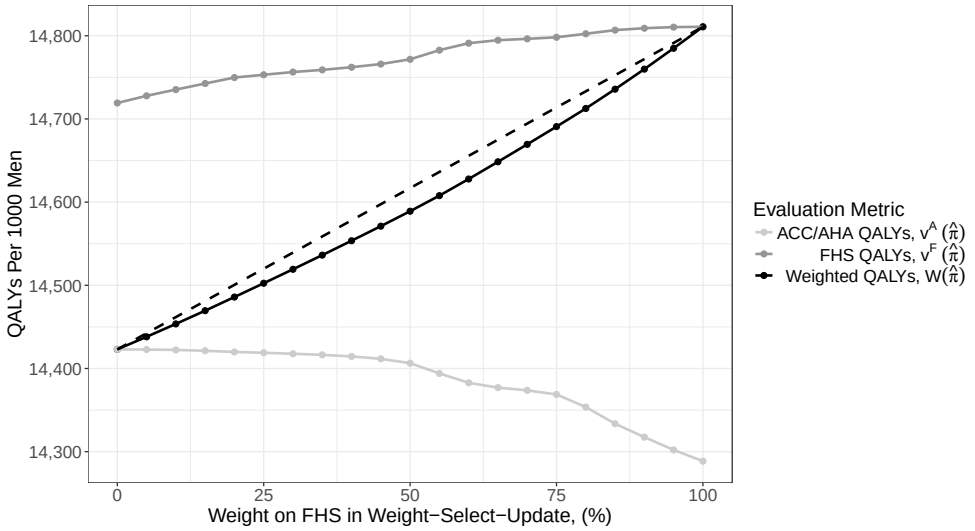
Using the MMDP described above, we evaluated the performance of the solutions generated via WSU in terms of computation time and the objective function of QALYs until first event. We also discuss the policy associated with the solution generated using WSU when the weights are treated as an uninformed prior on the models. The MMDP had 4099 states, 64 actions, 20 decision epochs, and 2 models.

Table 2.3: The solution times for the CVD MMDP. We report the time required to approximate a solution to the weighted problem using the Weight-Select-Update (WSU) algorithm and to solve each of the nominal models using standard dynamic programming, in CPU seconds.

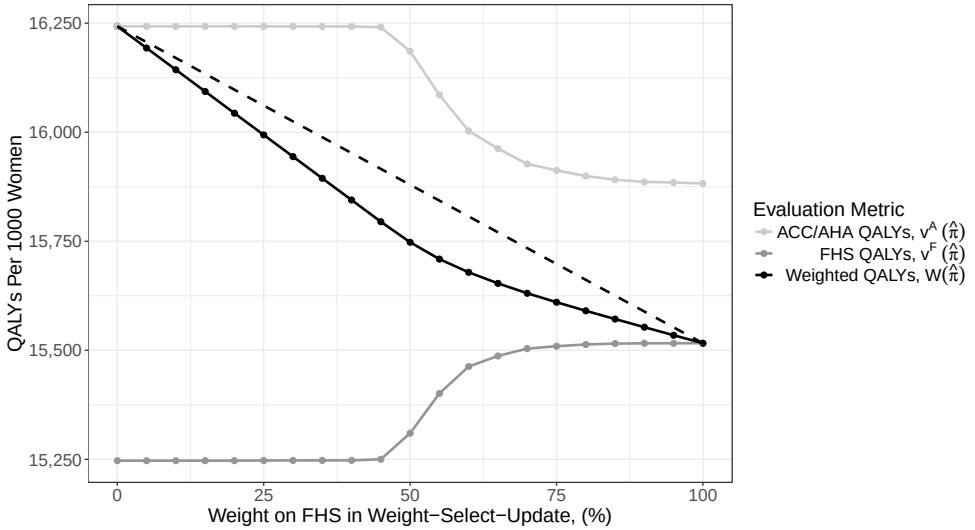
	Female	**Male**
WSU with $\lambda_F = \lambda_A = 0.5$	10.98 sec.	11.08 sec.
Standard DP, FHS Model	8.70 sec.	8.77 sec.
Standard DP, ACC/AHA Model	8.98 sec.	9.00 sec.

Table 2.3 shows the computation time required to run WSU with $\lambda_F = \lambda_A = 0.5$, as well as the time to required to solve the FHS model and the ACC/AHA model using standard dynamic programming, for the female and male problem parameters. While WSU requires more computation time than standard dynamic programming for each of the individual models, WSU does not take more computation time than the total time for solving both of the nominal models.

Figure 2.9 shows the performance of the policies generated using WSU when evaluated in the ACC/AHA and FHS models, as well as the weighted value of these two models for the corresponding choice of the weight on the FHS model, λ_F.The dashed line in these figures represents the upper bound from Proposition 2.8. When $\lambda_F = 100\%$, WSU finds the optimal policy for the FHS model which is why the maximum the FHS value is



(a) Male



(b) Female

Figure 2.9: The performance of the policies generated using the Weight-Select-Update (WSU) approximation algorithm in the CVD MMDP. The performance is reported for the MMDP for treatment of men (Figure 2.9a) and women (Figure 2.9b). For each choice of the weight on the FHS model in WSU, the graph shows the performance of these policies with respect to three different metrics: the performance in the ACC/AHA model (light grey), the performance in the FHS model (dark grey), and the weighted value (black). The dashed line represents the upper bound from Proposition 2.8.

achieved at $\lambda_F = 100\%$. Of the WSU policies, the worst value in the ACC/AHA model is achieved at this point because the algorithm ignores the performance in the ACC/AHA model. Analogously, when $\lambda_F = 0\%$, WSU finds the optimal policy for the ACC/AHA model which is why the performance in the ACC/AHA model achieves its maximum, and the performance in the FHS model is at its lowest value at this point. For values of $\lambda_F \in (0, 1)$, WSU generates policies that trade-off the performance between these two models. We found that WSU generated policies that slightly outperformed the policy generated by solving theMVP. As supported by Proposition 2.10, WSU has the desirable property that the performance in model m is non-decreasing in λ_m. For women, using the FHS model's optimal policy leads to a severe degradation in performance with respect to the ACC/AHA model. In contrast, WSU is able to generate policies that do not sacrifice too much performance in the ACC/AHA model in order to improve performance in the FHS model. The results for women clearly illustrate why taking a max-min approach instead of the MMDP approach can be problematic in some cases. To see this, note that the FHS model's optimal policy is a solution to the max-min problem because $v^F(\pi_F^*) < v^A(\pi_F^*)$ and thus no policy will be able to achieve a better value than π_F^* in the FHS model. However, Figure 2.9(b) shows that this policy leads to a significant degradation in performance in the ACC/AHA model relative to that model's optimal policy π_A^*. This demonstrates why taking a max-min approach, which is common in the robust MDP literature as pointed out in Section 2.2, can have the unintended consequence of ignoring the performance of a policy in all but one model in some cases. By taking the weighted value approach with nontrivial weights on the models, the DM is forced to consider the performance in all models. By generating policies using WSU by varying $\lambda_F \in (0, 1)$, the DM can strike a balance between the performance in the ACC/AHA model and the FHS model.

Table 2.4 illustrates that the WSU approximation algorithm generates a policy that will perform well in both the ACC/AHA model and in the FHS model. The table reports the QALYs gained per 1000 persons relative to a benchmark policy of never initiating treatment; these values are reported for three policies: (1) the ACC/AHA model's optimal policy, (2) the FHS model's optimal policy, and (3) the WSU policy. While using a model's optimal policy results in the highest possible QALY gain in that model, that model's optimal policy can sacrifice performance when evaluated in the other model. This is illustrated in the table in terms of *regret*; regret for a specific model is defined to be the difference between the QALYs gained by that model's optimal policy and the QALYs gained by the specified policy. The table shows that in the ACC/AHA model, the FHS model's optimal

(a) Male

Metric (per 1000 men)	Evaluation	ACC/AHA Optimal Policy	FHS Optimal Policy	WSU Policy
QALYs Gained Over No Treatment	ACC/AHA	695.9	561.5	679.3
	FHS	1788.9	1880.5	1841.4
	Weighted	1242.4	1211.0	1260.4
Regret	ACC/AHA	0	134.4	16.6
	FHS	91.6	0	39.1
	Weighted	45.8	67.2	27.9

(b) Female

Metric (per 1000 women)	Evaluation	ACC/AHA Optimal Policy	FHS Optimal Policy	WSU Policy
QALYs Gained Over No Treatment	ACC/AHA	205.2	−155.3	147.9
	FHS	1401.1	1670.4	1464.1
	Weighted	803.1	757.5	806.0
Regret	ACC/AHA	0	360.5	57.3
	FHS	269.3	0	206.3
	Weighted	134.7	180.2	131.8

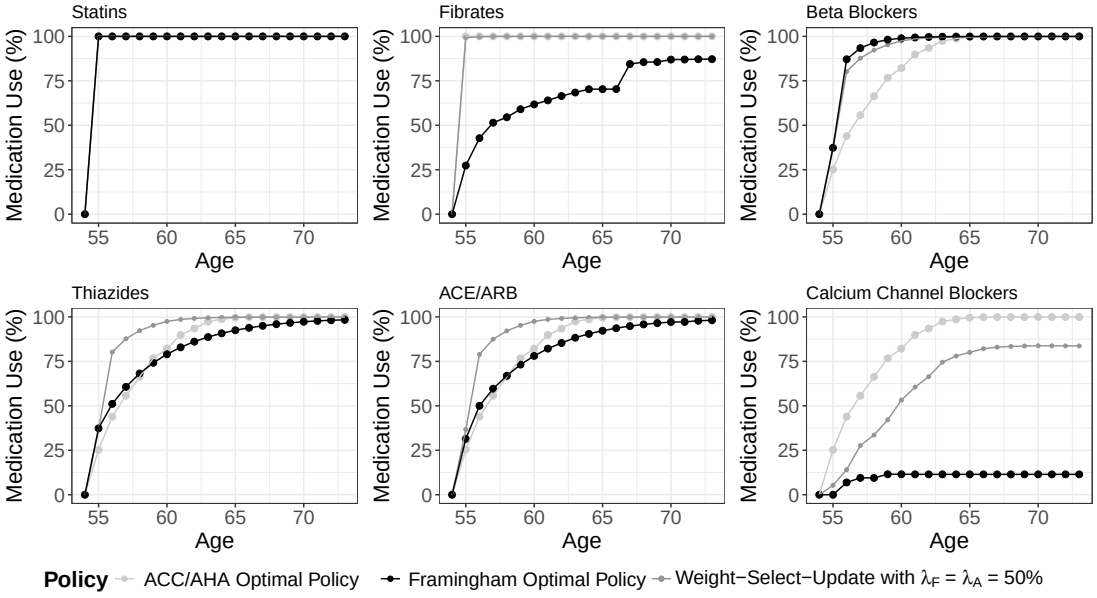
Table 2.4: A comparison of WSU and nominal policies for CVD MMDP. We report the performance of 3 policies in terms of QALYs gained over no treatment and regret for (a) men and (b) women. The 3 policies are (1) the optimal policy for the ACC/AHA model, (2) the optimal policy for the FHS model, and (3) the policy generated via the Weight-Select-Update (WSU) approximation algorithm which considers both the ACC/AHA and FHS models simultaneously. These policies are evaluated in terms of the QALYs gained over a policy which never initiates medication in the ACC/AHA model and the FHS model, as well as the weighted QALYs gained over no treatment in these two models. Regret is determined by taking the difference between the QALYs obtained by the optimal policy for a model and the QALYs obtained by the given policy.

policy achieves 134.4 QALYs per 1000 men less than the ACC/AHA model's optimal policy while the WSU policy is able to achieve only 16.6 less QALYs per 1000 men. Similarly, in the FHS model, the ACC/AHA model's optimal policy sacrifices 91.6 fewer QALYs per 1000 men relative to the optimal policy for the ACC/AHA model while the WSU policy only sacrifices 39.1 QALYs per 1000 men relative to the optimal policy for this model. Assuming an uninformed prior, the WSU approximation algorithm with equal weights on the models provides a weighted regret that is 17.9 and 2.9 QALYs less than the ACC/AHA model's optimal policy for men and women respectively, and WSU achieved a weighted regret that was 39.3 and 48.4 QALYs less than the FHS models' optimal policy for men

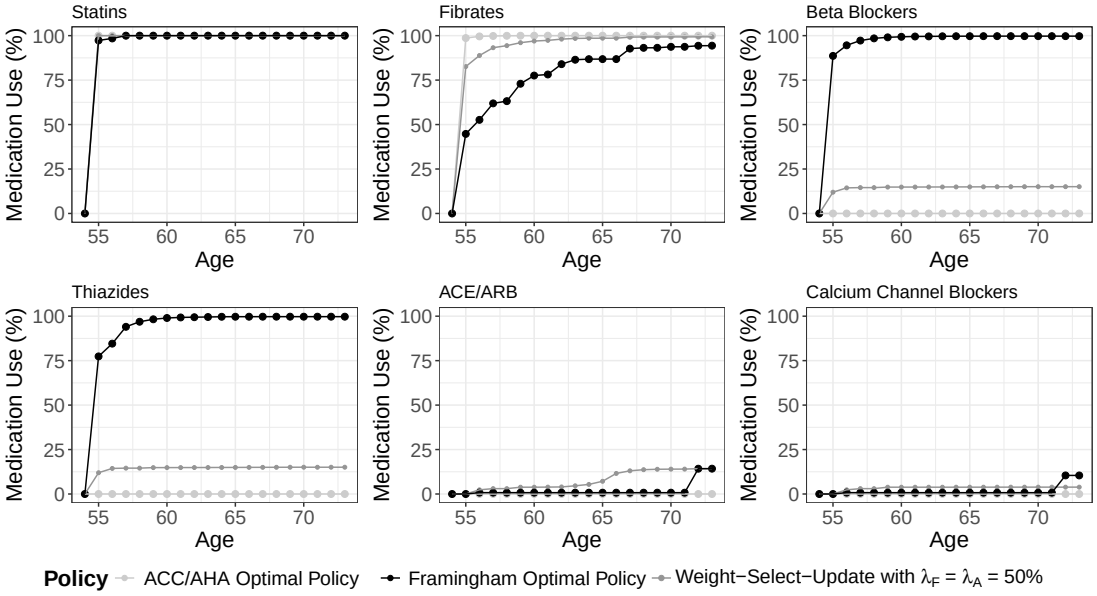
and women respectively. For women in particular, we find that using ignoring ambiguity in the risk calculations could potentially lead to very poor outcomes. The findings suggest that the FHS model's optimal policy is worse than the no treatment policy in the ACC/AHA model results. This is likely because the FHS model's optimal policy is much more aggressive in terms of starting medications (as seen in Figure 2.10, which is discussed later). Therefore, it seems that the FHS model's optimal policy is starting many women on medication which leads them to incur the disutility associated with these medications, but that these medications do not provide much benefit in terms of risk reduction in the ACC/AHA model. While the ACC/AHA model's optimal policy outperforms the no treatment policy in the Framingham model, we still see a large amount of regret in terms of QALYs gained per 1000 women in the FHS model. For both of these models, the WSU policy finds a policy that achieves a lower regret than the "other" model's optimal policy. Once again, weighting the regret from the two models equally, we see that the WSU policy is able to hedge against the ambiguity in risk for women and outperforms the two policies which ignore ambiguity.

It is interesting to note that the regret achieved by the WSU is much smaller for men than for women. This may be due to the disparity in the effects of ambiguity on decision making for women and men. EVPI is one way to quantify the expected value of resolving ambiguity and gives a DM a sense of how valuable it would be to obtain better information. Because $WSU \leq W^*$, the following is an upper bound on EVPI: $EVPI = WS - W^* \leq WS - WSU$. For this case study, the upper bound on the EVPI suggests that as many as 28 QALYs per 1000 men and 131.8 QALYs per 1000 women could be saved if there were no ambiguity in the cardiovascular risk of the patient. Estimates such as this provide insight into the value of future studies that could reduce the ambiguity.

Figures 2.10(a) and 2.10(b) illustrate medication use for male and female patients, respectively, under three different policies: the ACC/AHA model's optimal policy, the FHS model's optimal policy, and a policy generated via WSU with $\lambda_F = \lambda_A = 50\%$. These figures illustrate the probability that a patient who follows the specified policy from age 54 will be on the corresponding medication, conditioned on the patient being alive, as a function of their age. For men, the optimal policy for FHS model and the optimal policy for the ACC/AHA model agree that all men should start statins immediately, which could be explained by the relatively low disutility and high risk reduction of statins in both models. However, the models disagree in the use of fibrates and the 4 classes of blood pressure medications. The optimal policy for the ACC/AHA model suggests that all men



(a) Male



(b) Female

Figure 2.10: The medication usage in the CVD management MMDP. The figure presents the percentage of patients who have not died or had an event by the specified age that will be on a medication under each of three different treatment policies: the ACC/AHA model's optimal policy, the FHS model's optimal policy, and a policy generated via WSU with $\lambda_F = 50\%$, as evaluated in the FHS model.

should start fibrates immediately, suggesting that cholesterol control is important in the ACC/AHA model. However, fibrates are less commonly prescribed under the FHS model's optimal policy with about two-thirds of men on this medication by age 65. The policy generated with WSU agrees with the ACC/AHA policy's more extensive use of fibrates which may suggest that focusing on cholesterol control could be a good strategy in both models. Among the blood pressure medications, there are some disagreements between the optimal policies of the two models, with the most distinct being for the use of calcium channel blockers. This is likely to be due to the relatively high disutility (from side effects of calcium channel blockers) and low risk reduction associated with this medication. In the ACC/AHA model, the risk reduction of calcium channel blockers is worth the disutility in many cases, but in the FHS model, there are few instances in which the disutility associated with this medication is worth the gain in QALYs. The policy generated with WSU generates a policy that strikes a balance between these two extremes. While the differences are not quite as extreme, WSU also generates a policy that balances the utilization of thiazides prescribed by each model's optimal policy. For the other classes of blood pressure medications, both models agree that these medications should be commonly used for men, but disagree in the prioritization of these medications. The ACC/AHA model tends to utilize these medications more at latter ages, while the FHS model starts more men on these medications early. Interestingly, WSU suggests that starting ACE/ARBs and beta blockers earlier is a good strategy in both models.

For women, the optimal policy for FHS and the optimal policy for ACC/AHA agree that all women should be on a statin by age 57. The models mostly agree that relatively few women should start taking ACE/ARBs or calcium channel blockers. These results are not surprising as statins have low disutility and high risk reduction in both models, making them an attractive medication to use to manage a patients cardiovascular risk, while calcium channel blockers and ACE/ARBs are the two medications with lowest expected risk reduction in both models. The models disagree in how to treat women with thiazides, beta blockers, and fibrates. Beta blockers and thiazides have a higher estimated risk reduction in the FHS model than in the ACC/AHA model, which may be why these medications are considered good candidates to use in the FHS model but not in the ACC/AHA model. WSU finds a middle ground between the use of thiazides and beta blockers in the two models, but suggests more use of ACE/ARBs for some women.

In summary, the results of this case study illustrate how the policy generated by WSU trades off performance in the ACC/AHA and FHS models. This information could be useful

for decision makers who are tasked with designing screening and treatment protocols in the face of conflicting information from the medical literature.

2.8. Conclusions

In this chapter, we addressed the following research questions: (1) how can we improve stochastic dynamic optimization methods to account for parameter ambiguity in MDPs? (2) how much benefit is there to mitigating the effects of ambiguity? To address the first question, we introduced the MMDP, which allows for multiple models of the reward and transition probability parameter and whose solution provides a policy that maximizes the weighted value across these models. Solution of the non-adaptive MMDP provides a policy that is no more complicated than the policy corresponding to a single-model MDP while having the robustness that comes from accounting for multiple models of the MDP parameters. Although our complexity results establish that the MMDP model is computationally intractable, our analysis shows there is promising structure that can be exploited to create exact methods and fast approximation algorithms for solving the MMDP.

To address the second research question, we established connections between concepts in stochastic programming and the MMDP that quantify the impact of ambiguity on an MDP. We showed that the non-adaptive problem can be viewed as a two-stage stochastic program in which the first-stage decisions correspond to the policy and the second-stage decisions correspond to the value-to-go in each model under the specified policy. This characterization provided insight into a formulation of the non-adaptive problem as an MIP corresponding to the deterministic equivalent problem of the aforementioned two-stage stochastic program. We showed the adaptive problem is a special case of a POMDP and described solution methods that exploit the structure of the belief space for computational gain.

We evaluated the performance of our solution methods using a large set of randomly-generated test instances and also an MMDP of blood pressure and cholesterol management for type 2 diabetes as a case study. The WSU approximation algorithm performed very well across the randomly-generated test cases while solution of the MVP had some instances with large optimality gaps indicating that simply averaging multiple models should be done with caution. These randomly-generated test instances also showed that there was

very little gain from adaptive optimization of policies over non-adaptive optimization for the problem instances considered.

In the case study, we solved an MMDP consisting of two models which were parameterized according to two well-established but conflicting studies from the medical literature which give rise to ambiguity in the cardiovascular risk of a patient. The WSU policy addresses this ambiguity by trading off performance between these two models and is able to achieve a lower expected regret than either of the policies that would be obtained by simply solving a model parameterized by one of the studies, as is typically done in practice currently. The case study also highlights how the MMDP can be used to estimate the benefit of mitigating parameter ambiguity arising from these conflicting studies. The EVPI in this case study suggests that gaining more information about cardiovascular risk could lead to a substantial increase in QALYs, with potentially more benefit to be gained from learning more about women's cardiovascular risk. For the most part, the policies generated via the WSU approximation algorithm found a balance between the medication usage in each of the models. However, for men, the WSU approximation algorithm suggested that more aggressive use of thiazides and ACE/ARBs would be allow for a better balance in performance in both models. For women, the WSU approximation algorithm generated a policy that is more aggressive in cholesterol control than the FHS model's optimal policy and more aggressive in blood pressure control than the ACC/AHA model's optimal policy.

In summary, the MMDP is a new approach for incorporating parameter ambiguity in MDPs. This approach allows DMs to explicitly trade off conflicting models of problem parameters to generate a policy that performs well with respect to each model while keeping the same level of complexity as each model's optimal policy. The MMDP may be a valuable approach in many application areas of MDPs, such as medicine, where multiple sources are available for parameterizing the model.

Chapter 3

Decomposition methods for solving Multi-model Markov decision processes

3.1. Introduction

As discussed in Chapter 2, we propose the MMDP as a method to design strategies that account for ambiguity in the parameters of an MDP [70]. The MMDP represents ambiguity through multiple models of the MDP parameters, and a solution of the MMDP is a strategy that maximizes a weighted function of the performance in each of the models. We showed that searching for such a strategy within the class of Markov deterministic policies can be viewed as a two-stage stochastic integer program in which each model of the data corresponds to a scenario. Binary first-stage decision variables encode the Markov deterministic policy, and continuous second-stage decision variables encode the value of this policy in each of the models. Viewing the problem through a stochastic programming lens suggests promising approaches for solution methods, such as a MIP representing the extensive form of the deterministic equivalent problem and a corresponding decomposition scheme. However, the problem is NP-hard, and as we found in Chapter 2, solving the MIP directly is only viable for small-scale instances of MMDPs.

In this chapter, we present new methods for solving the MMDP that leverage the special structure of the problem. Specifically, we present a branch-and-cut (B&C) method which follows from decomposition algorithms in the stochastic programming literature and a custom branch-and-bound (B&B) method that exploits the decomposable nature of the problem. Further, we present the first numerical study of exact algorithms for MMDPs for realistic problem instances. Our numerical experiments show that the custom B&B

approach outperforms a commercial solver when applied to both the extensive form of the MIP and the customized B&C procedure in terms of computation time. Due to this new ability to solve larger MMDPs, we investigate the impact of ambiguity on the performance of an MDP in the context of a case study related to the optimal time to repair a machine. Our numerical results uncover some important properties of MMDPs and show that the MMDP approach is most beneficial when there is high variance in the transition probabilities given by the different models of the MDP.

The remainder of this chapter is structured as follows: In Section 3.2, we provide background on ambiguity in MDPs and related solution methods for solving MMDPs. In Section 3.3, we state the MIP for solving the WVP for a non-adaptive MMDP. In Section 3.4, we describe two methods, a B&C procedure and a B&B procedure, that leverage problem structure to solve MMDPs. We compare the solution methods in Section 3.5 using a machine maintenance problem. Finally, we conclude with a summary and discussion of the most important contributions of this chapter in Section 3.6.

3.2. Background

As discussed in Chapter 2, the MMDP is an approach for addressing ambiguity in MDPs in which the DM considers multiple models of the MDP parameters and seeks to find a policy that maximizes the weighted performance with respect to each model of the MDP. The MMDP is an NP-hard problem that has close ties to two-stage stochastic integer programs. In a two-stage stochastic program context, ambiguous problem parameters are treated as random variables where collective outcomes define *scenarios*. The DM must take some *first-stage* decisions before these random variables are realized. Upon the realization of the problem parameters, the DM may take some *recourse* actions in the *second-stage* to adapt to the information gained. In many cases, the random variables representing the problem data have finite support and the outcomes of these random variables are referred to as *scenarios*. We refer the reader to Birge and Louveaux [9] and Shapiro, Dentcheva, and Ruszczyski [63] for more information on stochastic programming.

Through the lens of stochastic programming, the MMDP can be viewed as a two-stage stochastic integer program in which the DM specifies a policy that is encoded through the use of first stage binary variables to indicate the action for each state-time pair. Then, the DM observes which model of the MDP is realized, which subsequently determines

the corresponding value functions for each model, which are represented using continuous variables. This two-stage stochastic integer program has a fixed *technology matrix* and a random *recourse matrix* with a finite number of scenarios. The problem can be written in its *extensive form* which contains decision variables representing the first-stage variables as well as second-stage decision variables for each scenario. However, the extensive form of the problem can become quite large and potentially inefficient to solve as the number of models grows. Fortunately, the constraint matrix in the extensive form is block-separable except for the columns corresponding to the first-stage decision variables. Therefore, fixing the first-stage "complicating" first-stage decision variables separates the problem into smaller, independent optimization problems that correspond to each scenario. Due to this structure, two-stage stochastic programs lend themselves well to divide-and-conquer type algorithms, such as Benders decomposition (also known as the L-shaped method) [5, 71]. Benders decomposition breaks the problem into a *master problem* and *subproblems*. The master problem typically only considers "complicating variables" while the subproblems will consider the other variables assuming fixed values of the complicating variables. In the context of stochastic programming, typically one solves a "relaxed master problem" which involves only the first-stage decisions and a subset of the constraints required to specify the complete optimization problem. Then, one uses duality for the subproblems to subsequently add constraints, or *cuts*, to the relaxed master problem to enforce the feasibility and optimality of the proposed first-stage solutions from the relaxed master problem.

The MMDP has some features that require special algorithmic consideration. First, the binary first-stage decision variables require integer programming methods, such as B&B. Early work to address this problem includes that of Wollmer [77], which proposed a cutting plane decomposition method, and Laporte and Louveaux [38] which proposed a B&C scheme wherein the feasibility and optimality cuts are added within a B&B framework for solving the master problem. Second, the MMDP has relatively complete recourse and therefore, feasibility cuts are not required. Third, logical constraints are required to enforce that the value functions in each of the models of the MMDP correctly correspond to the policy encoded by the binary variables. The logical constraints are enforced through the introduction of notorious "big-Ms" which weaken the linear programming relaxation of the extensive form of the MIP and cause problems for potential decomposition methods. Logic-based cuts have been proposed to strengthen formulations and avoid the explicit use of the big-M values [16, 34, 43, 53]. In this article, we propose a customized B&B method that eliminates the need to consider the big-Ms in the MMDP and further exploits the

unique structural properties of the MDP subproblems.

3.3. Model Formulation

In Chapter 2, we showed that the WVP for the Markov deterministic policy class, Π^{MD}, for an MMDP is a hard problem and propose the following MIP formulation in (3.1) as an exact solution method for the MMDP. For ease of reading, we restate the formulation below. The MIP formulation extends the standard LP formulation of an MDP [56, §6.9] to include continuous variables representing the value function for each model of the MMDPs and modifies the epigraph constraints to enforce that each model is evaluated according to the same policy. We define model-specific value function continuous variables such that $v_t^m(s) \in \mathbb{R}$ represents the value-to-go from state s at decision epoch t in model m. We define the *policy decision variables*

$$\pi_t(a|s) := \begin{cases} 1 & \text{if the policy states to take action } a \text{ in state } s \text{ at decision epoch } t, \\ 0 & \text{otherwise} \end{cases},$$

$\forall a \in \mathcal{A}, s \in \mathcal{S}, t \in \mathcal{T}$. Throughout this thesis, we will refer to (3.1) as the *extensive form* of the MMDP.

$$\max_{\pi,\, v} \quad \sum_{m\in\mathcal{M}} \sum_{s\in\mathcal{S}} \lambda_m \mu_1^m(s) v_1^m(s) \tag{3.1a}$$

$$\text{s.t.} \quad \sum_{a\in\mathcal{A}} \pi_t(a|s) = 1, \qquad \forall s \in \mathcal{S}, t \in \mathcal{T}, \tag{3.1b}$$

$$M\pi_t(a|s) + v_t^m(s) - \sum_{s'\in\mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \le r_t^m(s,a) + M, \forall m \in \mathcal{M}, s \in \mathcal{S}, \tag{3.1c}$$

$$a \in \mathcal{A}, t \in \mathcal{T},$$

$$v_{T+1}^m(s) \le r_{T+1}^m(s), \qquad \forall m \in \mathcal{M}, s \in \mathcal{S}, \tag{3.1d}$$

$$\pi_t(a|s) \in \{0,1\}, \qquad \forall a \in \mathcal{A}, s \in \mathcal{S}, t \in \mathcal{T}. \tag{3.1e}$$

Constraints (2.13b) and (2.13e) are used to encode a valid Markov deterministic policy. Constraint (2.13c) is a logic-based constraint which uses "big-M"s to enforce the relation-

ship between the value functions in each model and the policy, so long as $M \in \mathbb{R}$ is selected sufficiently large enough. Together, constraints (2.13c) and (2.13d) ensure that the value functions in each model correspond to the policy encoded via the binary π variables. As mentioned previously, the MMDP can be viewed as a two-stage stochastic integer program with binary first-stage decision variables, continuous second-stage decision variables, and relatively complete random recourse. Given a fixed policy, the MMDP reduces to evaluation of $|\mathcal{M}|$ Markov chains. However, policy optimization for an MMDP is challenging due to the coupling constraints that enforce the same policy is used in each model in the MMDP.

3.4. Methods that leverage problem structure

In this section, we describe two methods that leverage special structure in the problem to solve MMDPs. The first is a B&C approach for solving the MMDP MIP which is in the spirit of Benders decomposition. The B&C procedure follows from the view of the problem as a two-stage stochastic program with binary first-stage variables and continuous second-stage variables. The second is a customized B&B approach which considers the problem as $|\mathcal{M}|$ independent MDPs with coupling constraints on the policy. Later, in Section 3.5, we compare these two methods to a generic branch-and-bound implementation using a commercial solver for an example problem in the context of deciding when to repair a deteriorating machine.

3.4.1. Branch-and-cut for the Multi-model Markov decision process

Our B&C approach uses a master problem that includes the binary policy variables and enforces constraints (3.1c) and (3.1d) via cutting planes that are generated from each model's subproblem. We begin by describing the decomposition of (3.1) into the master problem and subproblems, and then we describe a B&C algorithm that uses this decomposition approach.

First, we describe the decomposition of (3.1) into a master problem and model-specific subproblems. The value of a fixed policy, π, in a model m is determined by solving a linear

program, which we refer to as Subproblem$^m(\pi)$:

$$\max_{v} \quad \sum_{s\in\mathcal{S}} \mu_1^m(s) v_1^m(s) \tag{3.2a}$$

$$\text{s.t.} \quad v_t^m(s) - \sum_{s'\in\mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \le r_t^m(s,a) + M - M\pi_t(s,a), \quad \forall s\in\mathcal{S}, a\in\mathcal{A}, t\in\mathcal{T}, \tag{3.2b}$$

$$v_{T+1}^m(s) \le r_{T+1}^m(s), \qquad \forall s\in\mathcal{S}. \tag{3.2c}$$

For constraints of the form (3.2b), we assign dual variables $x_t^m(s,a)\in\mathbb{R}$ and for constraints of the form (3.2c) we assign dual variables $x_{T+1}^m(s)\in\mathbb{R}$. Then, the dual of Subproblem$^m(\pi)$ is:

$$\min_{x} \quad \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t\in\mathcal{T}} \big(r_t^m(s,a) + M(1-\pi_t(a|s))\big) x_t^m(s,a) + \sum_{s\in\mathcal{S}} r_{T+1}(s) x_{T+1}^m(s) \tag{3.3a}$$

$$\text{s.t.} \quad \sum_{a\in\mathcal{A}} x_1^m(s,a) = \mu_1^m(s), \quad \forall s\in\mathcal{S}, \tag{3.3b}$$

$$\sum_{a\in\mathcal{A}} x_t^m(s,a) - \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} p_{t-1}^m(s|s',a) x_{t-1}^m(s',a') = 0, \qquad \forall s\in\mathcal{S}, t\in\mathcal{T}, \tag{3.3c}$$

$$x_{T+1}(s) - \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} p_T^m(s|s',a') x_T^m(s',a') = 0, \qquad \forall s\in\mathcal{S}, \tag{3.3d}$$

$$x_t^m(s,a) \ge 0, \qquad \forall s\in\mathcal{S}, a\in\mathcal{A}, t\in\mathcal{T}, \tag{3.3e}$$

$$x_{T+1}^m(s) \ge 0, \qquad \forall s\in\mathcal{S}. \tag{3.3f}$$

Given the policy, π, (3.3) is easy to solve because the constraint corresponding to the tuple (s,a,t) is binding if and only if $\pi_t(a|s) = 1$ so long as M is selected sufficiently large enough to enforce the logical relationship between the value functions and the policy. Therefore, given a policy $\pi_t(a|s)$, we have already identified an optimal basis for the subproblems. Because the primal constraint corresponding to (s,a,t) is non-binding if $\pi_t(a|s) = 0$, it follows from complementary slackness that $\pi_t(a|s) = 0 \Rightarrow x_t^m(s,a) = 0$.

Proposition 3.1. *The following forward substitution gives an optimal solution to* (3.3)*:*

$$x_1^m(s,a) = \begin{cases} \mu_1^m(s) & \text{if } \pi_1(a|s) = 1, \\ 0 & \text{otherwise}, \end{cases} \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}, \tag{3.4}$$

$$x_t^m(s,a) = \begin{cases} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{t-1}^m(s|s',a') x_{t-1}^m(s',a'), & \text{if } \pi_t(a|s) = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{3.5}$$

$$\forall s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T} \setminus \{1\}, \tag{3.6}$$

$$x_{T+1}^m(s) = \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_T(s|s',a') x_T^m(s',a'), \qquad \forall s \in \mathcal{S}. \tag{3.7}$$

Proof. First, we show that x as defined in (3.4)-(3.7) is feasible for (3.3). The solution generated by equation set (3.4)-(3.7) satisfies (3.3b) because

$$\sum_{a \in \mathcal{A}} x_1^m(s,a) = \sum_{a \in \mathcal{A}} \mu_1^m(s) \pi_1(a|s) = \mu_1^m(s) \sum_{a \in \mathcal{A}} \pi_1(a|s) = \mu_1^m(s), \quad \forall s \in \mathcal{S}. \tag{3.8}$$

The solution satisfies (3.3c) because

$$\sum_{a \in \mathcal{A}} x_t^m(s,a) = \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{t-1}^m(s|s',a') x_{t-1}^m(s',a') \pi_t(a|s) \tag{3.9}$$

$$= \sum_{a \in \mathcal{A}} \pi_t(a|s) \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{t-1}^m(s|s',a') x_{t-1}^m(s',a') \tag{3.10}$$

$$= \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{t-1}^m(s|s',a') x_{t-1}^m(s',a'), \forall s \in \mathcal{S}, t \in \mathcal{T} \setminus \{1\}, \tag{3.11}$$

and (3.3d) is satisfied by definition. The non-negativity constraints are also satisfied.

Next, we show that x as defined in (3.4)-(3.7) is optimal. Consider the following feasible solution to Subproblem(π):

$$v_{T+1}^m(s) = r_T^m(s), \quad \forall s \in \mathcal{S}, \tag{3.12}$$

$$v_t^m(s) = \begin{cases} r_t^m(s,a) + \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s) & \text{if } \pi_t(a|s) = 1 \\ 0 & \text{otherwise.} \end{cases}, \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T}. \tag{3.13}$$

The solutions v and x are feasible for Subproblem(π) and its dual. For all $(s,a,t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{T}$ such that $\pi_t(a|s) = 0$, $x_t^m(s,a) = 0$, and for all $(s,a,t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{T}$ such that $\pi_t(a|s) = 1$, $v_t^m(s) - \sum_{s' \in \mathcal{S}} v_{t+1}^m(s)) = r_t^m(s,a) + M(1 - \pi_t(a|s))$ which implies

$$x_t^m(s,a)\left(v_t^m(s) - r_t^m(s,a) - M + M\pi_t(a|s) - \sum_{s' \in \mathcal{S}} p_t(s'|s,a) v_{t+1}^m(s)\right) = 0, \qquad \forall s \in \mathcal{S},$$

$$a \in \mathcal{A}, t \in \mathcal{T},$$

$$x_{T+1}^m(s)\left(v_{T+1}^m(s) - r_{T+1}^m(s)\right) = 0, \qquad \forall s \in \mathcal{S},$$

$$v_1^m(s)\left(\sum_{a \in \mathcal{A}} x_1^m(s,a) - \mu_1^m(s)\right) = 0, \qquad \forall s \in \mathcal{S},$$

$$v_t^m(s)\left(x_t^m(s,a) - \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{t-1}(s|s',a') x_{t-1}^m(s',a')\right) = 0, \qquad \forall s \in \mathcal{S},$$

$$t \in \mathcal{T} \setminus \{1\},$$

$$v_{T+1}^m(s)\left(x_{T+1}^m(s) - \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_T^m(s|s',a') x_T^m(s',a')\right) = 0, \qquad \forall s \in \mathcal{S}.$$

Thus, by complementary slackness, v and x are optimal solutions for their respective problems. □

Corollary 3.1. (3.3) *can be solved in* $O(T|\mathcal{S}|^2)$ *time.*

Proof. Computing the values in (3.4) can be completed in $O(|\mathcal{S}|)$ time. For a given state, (3.6) can be completed in $O(|S|)$ time because exactly one action is taken in each state. Therefore, for a given decision epoch, the solution of (3.6) requires $O(|S|^2)$ time and thus, the total time required for substitutions in (3.6) can be completed in $O(T|S|^2)$ time. Equation set (3.7) also requires $O(|S|^2)$ time. □

For an optimal policy π^* for a given subproblem, the dual solution has a corresponding objective value of

$$z_m := \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}} \left(r_t^m(s,a) + M(1 - \pi_t^*(a|s))\right) x_t^m(s,a) + \sum_{s \in \mathcal{S}} r_{T+1}(s) x_{T+1}^m(s). \tag{3.14}$$

Therefore, z_m is a tight upper bound on the value that can be obtained by using the fixed policy π^* in model m. Further, the hyperplane in (3.14) can be used to bound the value of other policies in model m.

To this point, we have described how the subproblem can be solved quickly for a fixed value of the policy π. We now describe the master problem (3.15) which is a MIP that uses binary variables, $\pi \in \{0,1\}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{T}|}$, to encode the policy and continuous variables, $\theta \in \mathbb{R}^m$, to be used as surrogate variables for the value functions in each model. The master problem incorporates optimality cuts generated from the subproblems corresponding to previously investigated policies, $\pi^k, k = 1, \ldots, K$.

$$
\begin{aligned}
&\max_{\pi,\theta} \quad \sum_{m\in\mathcal{M}} \lambda_m \theta_m \\
&\text{s.t.} \\
&\sum_{a\in\mathcal{A}} \pi_t(a|s) = 1, && \forall s \in \mathcal{S}, t \in \mathcal{T}, \\
&\theta_m \leq \sum_{(s,a,t)\in\mathcal{S}\times\mathcal{A}\times\mathcal{T}} x_k^m(s,a,t)\big(r_t^m(s,a) + M - M\pi_t(a|s)\big) && \forall s \in \mathcal{S}, a \in \mathcal{A}, \\
&\qquad + \sum_{s\in\mathcal{S}} r_{T+1}^m(s) x_k^m(s, T+1), && t \in \mathcal{T}, k = 1, \ldots, K, \\
&v_{T+1}^m(s) \leq r_{T+1}^m(s), && \forall s \in \mathcal{S}, \\
&\pi_t(a|s) \in \{0,1\}, && \forall a \in \mathcal{A}, s \in \mathcal{S}, t \in \mathcal{T}
\end{aligned}
\tag{3.15}
$$

Algorithm 4 is a B&C algorithm for solving the MMDP that uses the decomposition described above. The algorithm we present largely follows from the Integer L-Shaped Method [38], but leverages the special structure of the subproblems to quickly generate optimality cuts.

3.4.2. Branch-and-bound for the Multi-model Markov decision process

In this section, we present a customized B&B framework that can be used to solve MMDPs. We begin by introducing the general framework of the MMDP B&B procedure, and then subsequently describe strategies for search, branching, and pruning within this framework.

Algorithm 4: Branch-and-cut for MMDP

Initialize Set $k := 0$, $\nu := 0$. Let $\bar{\pi}$ be any feasible policy and $\bar{z}$ be the corresponding weighted value

Select a pending node from the list; if none exists, stop.

Set $\nu := \nu + 1$; Solve the current instance of problem (3.15).

if *Current problem has no feasible solution* **then**

  Fathom the current node; Return to Step 2

else

  Let (π^ν, θ^ν) be an optimal solution to the current problem.

end

if $\theta^\nu < \bar{z}$ **then**

  Fathom the current node. Return to Step 2

end

if *The current solution* π^ν *violates an integer constraint* **then**

  Create two new pendant nodes; Return to Step 2

end

Use (3.4)-(3.7) to obtain the dual solution $x_m^\nu(s,a,t) := x_t^m(s,a), x_m^\nu(s,T+1) := x_{T+1}^m(s),\ \forall m \in \mathcal{M}$

Let z_m^ν be the corresponding objective function value as in (3.14) and $z^\nu = \sum_{m\in\mathcal{M}} \lambda_m z_m^\nu$

if $z^\nu > \bar{z}$ **then**

  Update the incumbent to be (π^ν, z^ν)

end

if $\theta^\nu \leq z^\nu$ **then**

  Fathom the current node. Return to Step 2

else

  For each model such that $(\theta^\nu > z^\nu)$, add an optimality cut to (3.15):

$$\theta_m \leq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{t\in\mathcal{T}} \big(r_t^m(s,a) + M(1-\pi_t(a|s)\big) x_m^\nu(s,a,t) + \sum_{s\in\mathcal{S}} r_{T+1}^m(s) x_m^\nu(s,T+1)$$

  Set $k = k+1$. Return to Step 3

end

General Branch-and-bound framework

The B&B procedure is grounded in the idea that if we relax the requirement that each of the $|\mathcal{M}|$ MDPs must have the same policy, the problem becomes easy to solve as it involves solving $|\mathcal{M}|$ independent MDPs. Thus, at the beginning of the B&B procedure, we completely relax the requirement at the *root node* that each of the $|\mathcal{M}|$ MDPs must have the same policy, and solve each individual model via backward induction to obtain an optimal policy for each model. Then, we sequentially add restrictions that the policies used in certain (s,t)-pairs must be the same across the $|\mathcal{M}|$ MDPs. We refer to these restrictions as *partial policies* because they specify the actions to be taken in some, but not necessarily all, of the (s,t)-pairs. Given a partial policy, one solves each MDP independently, taking the best action for each (s,t)-pair in that given model unless that (s,t)-pair's action has already been fixed in the partial policy.

The B&B procedure begins with an empty partial policy, meaning that none of the (s,t)-pairs are fixed to a specific action, and thus, each of the $|\mathcal{M}|$ MDPs can be solved independently. For a given partial policy, $\hat{\pi}$, the corresponding relaxation of the MMDP is solved. The solution of the relaxation will be described in more detail in Section 14. Solving the relaxation provides an optimistic completion of the partial policy of a given node and an upper bound on the weighted value objective that could be achieved by completing this partial policy. The upper bound is compared to the value associated with the *incumbent*, the best solution seen so far. If the upper bound is worse than the lower bound associated with the incumbent, the node is pruned, meaning that none of its descendents in the tree will be examined. If the model-specific optimal completions of the partial policy are the same in each of the $|\mathcal{M}|$ MDPs then the policy is an *implementable policy* for the MMDP and, if the corresponding weighted value is better than the incumbent, then the incumbent solution is replaced with this policy and the node is pruned by optimality. If the partial policy is not pruned, the node is added to the list of pending nodes. At each iteration a pending node (i.e. partial policy) is selected for branching.

Proposition 3.2 (Worst-case running time of the B&B procedure)**.** *The worst-case running time of the B&B procedure is* $O(|\mathcal{M}|TS^2A^{ST+1})$ *where* $T = |\mathcal{T}|$, $S = |\mathcal{S}|$, *and* $A = |\mathcal{A}|$.

Proof. B&B algorithms have a worst-case running time of $O(Ub^d)$ where b is the branching factor, d is the depth of the tree, and U is an upper bound on time required to solve a subproblem [11]. The branching factor of the B&B tree described above is A and the

Algorithm 5: Solve Relaxation($\hat{\pi}^i$) for the MMDP

Data: Partial policy $\hat{\pi}^i$
$v^m_{T+1}(s) \leftarrow r^m_{T+1}(s), \forall s \in \mathcal{S}, m \in \mathcal{M}$
$t \leftarrow T$
while $t > 0$ **do**
  for $(s, m) \in \mathcal{S} \times \mathcal{M}$ **do**
    if $\hat{\pi}^i_t(s)$ *fixed* **then**
      $v^m_t(s) \leftarrow r^m_t(s, \hat{\pi}^i_t(s)) + \sum_{s' \in \mathcal{S}} p^m_t(s'|s, \hat{\pi}^i_t(s)) v^m_{t+1}(s')$
    else
      $v^m_t(s) \leftarrow \max_{a \in \mathcal{A}} \left\{ r^m_t(s, a) + \sum_{s' \in \mathcal{S}} p^m_t(s'|s, a) v^m_{t+1}(s') \right\}$
      $\pi^m(s, t, \hat{\pi}^i) \leftarrow arg\,max_{a \in \mathcal{A}} \left\{ r^m_t(s, a) + \sum_{s' \in \mathcal{S}} p^m_t(s'|s, a) v^m_{t+1}(s') \right\}$
    end
  end
end
$v^m(\hat{\pi}^i) \leftarrow \sum_{s \in \mathcal{S}} \mu^m_1(s) v^m_1(s), \quad \forall m \in \mathcal{M}$
$\bar{z}^i \leftarrow \sum_{m \in \mathcal{M}} \lambda_m v^m(\hat{\pi}^i)$

depth of the tree is ST. The worst-case running time to solve a subproblem is the time required to solve M MDPs in the worst-case. The worst-case time to solve a single MDP is $O(TS^2A)$ [56, p. 93]. □

Interestingly the worst-case running time grows linearly in the number of models considered and exponentially in the number of actions. This suggests that identifying actions than cannot be optimal via an action elimination procedure [56, §6.7.2] or the presence of special structure such as a monotone optimal policy [56, §6.11.2] could be used to reduce computation time.

Search strategy

Standard options for the search strategy include breadth-first search (BrFS), depth-first search (DFS), and best-first search (BFS). In DFS, the unexplored partial policies are explored in last-in-first-out fashion, such that there is a priority placed on finding implementable policies. In this search strategy, the list of partial policies is maintained using a stack structure wherein the algorithm explores the partial policy of the most recently child node first before generating any of its siblings. One advantage of this approach is that often there are some value function estimates corresponding to a partial policy that can be reused by the children of that partial policy. In other types of searches, the reusable value

function estimates would need to be stored while awaiting selection of the node, however in DFS this information can be removed from memory as soon as the policy is optimally completed. There can be drawbacks to DFS. For example, DFS can lead to imbalanced search trees in which some partial policies remain pending at small depths in the tree while the algorithm is focused on completing other partial policies, which could lead to the B&B procedure spending a lot of time completing poor policies before finding an optimal policy.

Another search strategy is BrFS which can be viewed as a first-in-first-out approach to exploring partial policies, meaning that all children of a node are examined before any grandchildren can be examined. In this case, the unexplored partial policies roughly have the same number of (s,t)-pairs fixed at each stage in the B&B procedure. However, there are usually substantial memory requirements associated with BrFS because most children nodes are not explored immediately after their parent nodes. Another search strategy is the BFS which considers the partial policies that appear most promising. In this setting, we use a best-bound approach which explores a partial policy $\hat{\pi}$ next if its corresponding upper bound $\hat{z}$ is higher than all the other unexplored policies.

Branching strategy

At each iteration in which at least one pending node exists, the B&B algorithm selects an (s,t)-pair to branch on to generate new subproblems. For the MMDP B&B procedure, we focus on *wide branching* strategies which generate $|\mathcal{A}(s,t)|$ new partial policies to be investigated upon branching on a pair (s,t), where each new subproblem corresponds to fixing $\pi_t(s)$ to be each of the possible actions in $\mathcal{A}(s,t)$, the action set specific to this $(s,t)-$pair. There are several strategies for branching on (s,t)-pairs.

One such branching strategy is *horizon-based* branching. In this strategy, decisions for branching are made on the basis of the decision epoch. That is, there is some ordering of the (s,t)-pairs such that $t < t'$ implies something about the order in which $\pi_t(s)$ and $\pi_{t'}(s')$ are fixed for any states s and s'. One such approach is *early-horizon branching* in which the decisions for epochs early in the planning horizon are fixed first. Early-horizon branching may be desirable because this allows the branch-and-bound procedure to reuse value function estimates from the *wait-and-see* problem wherein each of the $|\mathcal{M}|$ MDPs is solved independently.

Another branching strategy is *disagreement branching*, which is in the vein of *most fractional* branching in integer programming. The idea behind the disagreement branching

strategy is to select the parts of the optimal completion of the partial policy found in the relaxation, and select the (s,t)-pair for which there is the largest disagreement among the models' optimal completion policies. For instance, one disagreement-based strategy is to select the (s,t)-pair for which there is the largest number of actions suggested by the different models.

Pruning strategy

A natural pruning procedure in this context is to eliminate partial policies based on their corresponding upper bounds. If the upper bound on the weighted value corresponding to the completion of partial policy $\hat{\pi}$ is lower than the weighted value associated with the incumbent solution, then the B&B procedure no longer needs to consider partial policy $\hat{\pi}$ as there is no way to complete the policy such that it will be better than the incumbent. We restrict our attention to this pruning strategy because the upper bound associated with a partial policy is easily obtained by solving the relaxation via Algorithm 5. When the partial policy is empty, this relaxation corresponds to solving the wait-and-see problem. For any other partial policy such that some (s,t)-pairs have been fixed, this procedure uses backward induction to find the optimal completion of the partial policy.

3.5. Case study: Machine maintenance

In this section, we consider a version of the machine maintenance problem similar to that presented in Delage and Mannor to illustrate the solution of the MMDP using the algorithms of the previous section. We consider the machine replacement problem with 6 states, 3 actions, and a uniform initial state distribution. States 1 to 6 represent the quality of the machine, with 1 being the highest quality and 6 being the lowest. There are 3 actions: *Do Nothing*, Repair Option 1, and Repair Option 2. Repair Option 1 provides lower improvement at lower cost and Repair Option 2 provides higher improvement at the higher cost. Figure 3.1 illustrates the expected values of the transition probabilities for actions "Do Nothing" (Action 0) and *Repair Option 1* (Action 1). The transition probabilities for *Repair Option 2* are similar to *Repair Option 1* but allow for the machine to potentially improve two states in between epochs. Performing Repair Options 1 and 2 incur cost penalties of 5 and 8, respectively. We assume that the cost to operate the machine is dependent on the quality of the machine and that when the machine reaches

the age of 7, it will be replaced completely. The planning horizon consists of 6 decision epochs, and we ignore the cost to replace the machine at time 7.

We consider the case in which the DM does not know the transition dynamics of the machine exactly, but has multiple estimates of what these transition probabilities may be. To generate instances of the transition probability estimates, we draw samples from a Dirichlet distribution which is used to conduct probabilistic sensitivity analyses on Markov chains [12]. The Dirichlet distribution used has parameters $(\alpha\bar{p}_1, \ldots, \alpha\bar{p}_{|\mathcal{S}|})$ for each row of the transition probability matrix where the base measure of the distribution, $(\bar{p}_1, \ldots, \bar{p}_{|\mathcal{S}|})$, corresponds to the mean values of the parameters of that row. The concentration parameter $\alpha \in \mathbb{R}$ is varied to consider the influence of the variation in the estimates on the impact of ambiguity. For low values of α, the transition probability estimates have higher variation in the transition probability estimates, while sampling from a Dirichlet distribution with higher values of α produces samples that are more closely concentrated around the mean values of the transition probabilities. To illustrate this, we show 100 samples of the Dirichlet distribution for $\alpha = 0.5,\ 1,\ 10$, and 20 in Figure 3.2. Note that we use the Dirichlet distribution to study the impact of the variance in the rows of the transition matrices among the models on the computational performance and the value of the MMDP solution; we are not assuming that the DM has any prior information about the distribution of these models.

We consider concentration parameters $\alpha \in \{0.5, 1, 10, 20\}$ and $|\mathcal{M}| \in \{10, 20, 30\}$. For each of these combinations of α and $|\mathcal{M}|$, we generate 20 instances of the corresponding MMDP. For each instance, we sampled the rows for each state and action independently from the corresponding Dirichlet distribution corresponding to that row. We assume the transition probabilities were stationary. We also assume the rewards are stationary and independent of the model.

To solve the machine repair problem, we implemented the extensive form of the MIP formulation and the B&C procedure described in Algorithm 4 using the commercial solver Gurobi Optimizer Version 8.0.1 in C++ using XCode. The optimality cuts were added as *lazy constraints* within user callbacks whenever an integer feasible solution was found. We implemented the extensive form using default settings in Gurobi, and used special ordered set (SOS) Type 1 constraints for both the extensive form and B&C implementations. We implemented the custom B&B procedure in C++ in Xcode using a priority queue data structure to manage the search tree and a custom node class to manage the information stored at each node in the tree. We specified an optimality gap of 0.01% for each of the

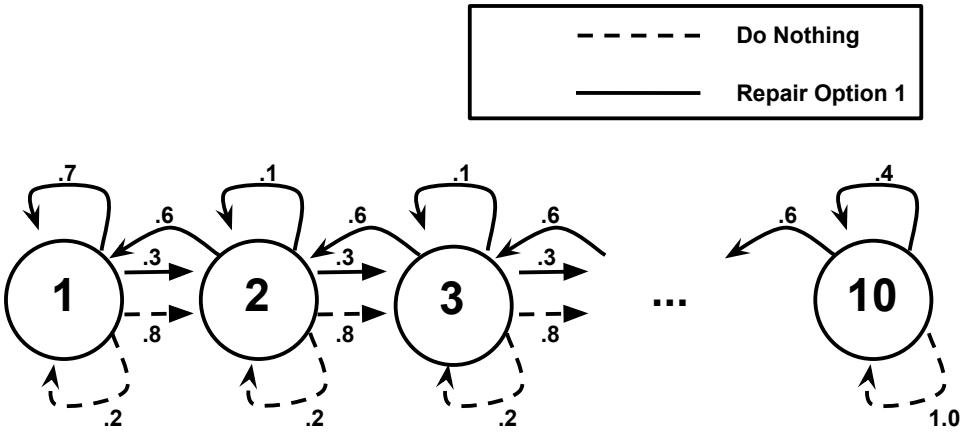
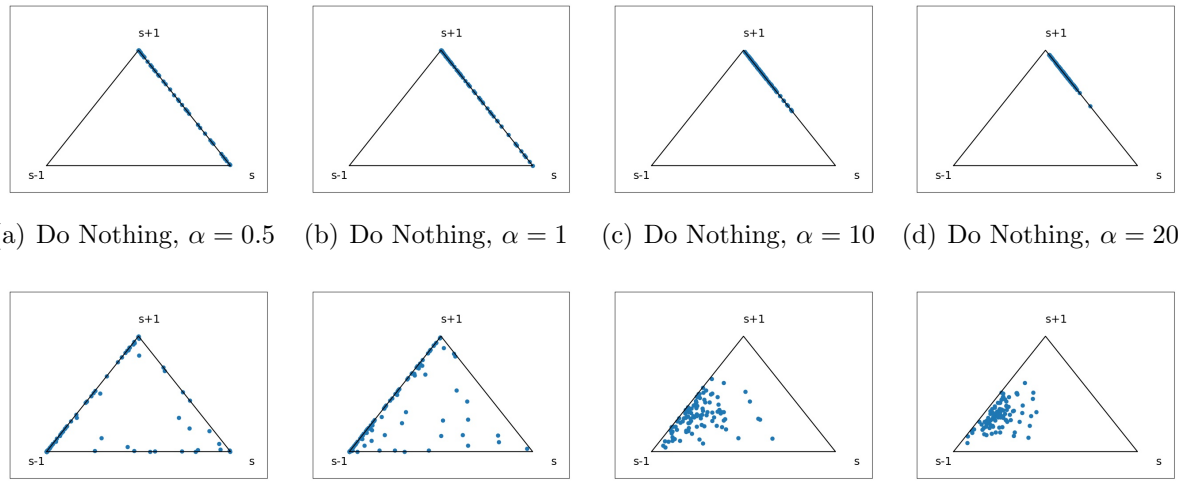


Figure 3.1: An illustration of the machine maintenance problem with one repair option. The states represent the quality of the machine. The dashed and solid lines represent stochastic transitions among these states corresponding to the "Do Nothing" action and the "Repair" action, respectively.

algorithms. For each solution method, we provided the solution from the approximation algorithm described in Chapter 2 as a *warm-start* in Gurobi for the extensive form and B&C, and as the incumbent for the B&B procedure.

All experiments were run using a single thread on a Macintosh MacBook Pro with 2.7 GHz CPUs and 16 GB of memory. A time limit of 300 seconds was enforced.

We compared the MMDP policy to the mean value policy which is determined by solving a single MDP obtained by averaging the estimates of the transition probabilities. We report two standard metrics for measuring the impact of ambiguity in two-stage stochastic programs: (1) the EVPI which measures the value of knowing the transition probabilities precisely before selecting the policy and (2) the VSS which represents the improvement in cost from considering each model of the MMDP rather than simply averaging the models and solving a single MDP. We report EVPI as a relative improvement over the cost obtained by the MMDP policy and we report the value of the MMDP policy as a relative improvement over the mean value policy.



(a) Do Nothing, $\alpha = 0.5$ (b) Do Nothing, $\alpha = 1$ (c) Do Nothing, $\alpha = 10$ (d) Do Nothing, $\alpha = 20$

(e) Repair, $\alpha = 0.5$ (f) Repair, $\alpha = 1$ (g) Repair, $\alpha = 10$ (h) Repair, $\alpha = 20$

Figure 3.2: Samples from the Dirichlet distributions used in the machine repair case study. Each figure represents 100 samples from the Dirichlet distributions for the corresponding action and *concentration parameter*, α, that are used to generate the models for the case study. The triangle represents the probability simplex defining the transition probabilities from state s to $s-1$ (lower left corner of the triangle), s (lower right corner) and $s+1$ (top corner). The top row 3.2(a)-3.2(d) corresponds to transitions for the action to "Do Nothing"; these probabilities were drawn from a Dirichlet distribution with a mean $(0, 0.2, 0.8)$. These figures represents samples from the Dirichlet distribution with parameters $(0, 0.2\alpha, 0.8\alpha)$ for $\alpha = 5,\ 10,\ 20$ and 30, respectively. The bottom row 3.2(e)-3.2(h) corresponds to transitions for the action to "Repair" with option 1; these probabilities were drawn from a Dirichlet distribution with a mean $(0.6, 0.1, 0.3)$. Each figure represents samples from the corresponding Dirichlet distribution with parameters $(0.6\alpha, 0.1\alpha, 0.3\alpha)$ for $\alpha = 5,\ 10,\ 20$ and 30, respectively.

Table 3.1: Measures of ambiguity and computational performance of the extensive form, Branch-and-cut, and Branch-and-bound on the machine maintenance MMDP for various values of the concentration parameter, α, and number of models $|\mathcal{M}|$. The warm start provided was the solution found using the approximation algorithm described in Chapter 2. A time limit of 300 seconds was enforced.

$\|\mathcal{M}\|$	α	Measures of Ambiguity				Solution Time (CPU Seconds)						Optimality Gap (%)					
		VSS(%)		EVPI (%)		Extensive Form		Branch-and-Cut		Branch-and-Bound		Extensive Form		Branch-and-Cut		Branch-and-Bound	
		Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.
10	0.5	3.69	7.46	19.55	25.19	>300.00	>300.00	>300.00	>300.00	14.87	80.76	19.65	25.19	22.38	26.98	<0.01	<0.01
	1.0	2.15	5.55	17.43	24.08	>300.00	>300.00	> 300.00	> 300.00	5.56	37.07	17.45	24.08	19.17	26.33	<0.01	<0.01
	10.0	0.03	0.11	4.56	7.86	>300.00	>300.00	>300.00	>300.00	0.11	0.62	4.56	7.86	4.58	7.86	<0.01	<0.01
	20.0	0.00	0.08	2.34	4.32	> 300.00	> 300.00	>300.00	> 300.00	0.04	0.35	2.34	4.32	2.34	4.40	<0.01	<0.01
20	0.5	3.36	6.66	22.82	29.65	>300.00	>300.00	>300.00	>300.00	70.07	291.78	23.08	30.29	25.31	34.33	<0.01	<0.01
	1.0	2.07	5.31	19.92	28.47	>300.00	> 300.00	>300.00	>300.00	48.44	> 300.00	19.91	25.60	21.39	27.40	0.18	3.55
	10.0	0.01	0.04	5.00	7.19	>300.00	> 300.00	>300.00	>300.00	0.21	0.59	5.00	7.19	5.00	7.19	<0.01	<0.01
	20.0	0.00	0.01	2.87	4.27	> 300.00	> 300.00	> 300.00	> 300.00	0.09	0.22	2.87	4.27	2.87	4.27	<0.01	<0.01
30	0.5	2.77	5.79	22.93	35.29	> 300.00	> 300.00	>300.00	> 300.00	87.40	> 300.00	22.96	28.90	24.72	32.87	0.35	6.86
	1.0	1.37	3.90	20.95	26.93	> 300.00	> 300.00	> 300.00	> 300.00	72.97	> 300.00	20.97	24.29	21.87	25.74	0.15	2.97
	10.0	0.01	0.07	5.15	7.93	> 300.00	> 300.00	> 300.00	> 300.00	0.28	1.06	5.16	7.93	5.16	7.93	<0.01	<0.01
	20.0	0.01	0.03	2.66	3.79	> 300.00	> 300.00	>300.00	> 300.00	0.10	0.14	2.66	3.79	2.67	3.79	<0.01	<0.01

Table 3.1 summarizes the results of our experiments. First, we consider the influence of the concentration parameter, α, and the number of models $|\mathcal{M}|$ on measures of ambiguity. We observe VSS decreases significantly as the concentration parameter, α, increases and as the number of models increases. As the concentration parameter increases, the variance in the transition probabilities across the models decreases. This suggests that if the DM has transition probability models that are closely concentrated around their mean, solving the MVP may provide a policy that performs well across the models. However, when the concentration parameter is higher, there is more ambiguity in the true dynamics of the system. Thus, when there is high variation in the models of the MDP, the MMDP tends to find policies that perform well with respect to the various models relative to the solution of a single MDP obtained by averaging the models' parameters. The EVPI is a measure of ambiguity that could be interpreted as the expected regret in each model obtained by implementing the MMDP policy instead of the optimal policy for that model. As with VSS, we see that EVPI decreases as the variance among the models decreases. The impact of the number of models on EVPI is less clear.

We also present the solution times and optimality gaps for the extensive form, B&C, and B&B solution methods. The extensive form and B&C solution procedures solved 0 of the 240 test instances, while B&B solved 235 of the 240 instances within the time limit. Even in the worst-case, the custom B&B procedure performed very well with an optimality gap of less than 7% in all problem instances. For the B&B procedure, we observe that, in general, the average time to solve the MMDP increases as the number of models increases and as the variance among the models increases. We see that solving the extensive form of the MMDP directly outperforms the B&C procedure in terms of optimality gap after the time limit.

3.6. Conclusions

In this chapter, we addressed the problem of solving large MMDPs. The MMDP has been proposed as a method for designing policies that account for ambiguity in the input data for MDPs, but the solution of MMDPs had been restricted to a small number of models. Finding an optimal Markov deterministic policy for an MMDP is an NP-hard problem, and the extensive form of the problem is a MIP that includes "big-M"s which weaken the linear programming relaxation.

We proposed two decomposition methods that leverage the problem structure to solve the MMDP. The first was a B&C algorithm in the vein of the Integer L-Shaped Method for two-stage stochastic integer programming with binary first-stage variables. The B&C algorithm decomposed the extensive form of the MIP into a master problem involving the binary variables used to encode the policy and $|\mathcal{M}|$ subproblems that evaluate a proposed policy in each model of the MDP. Unfortunately, the extensive form relied on the notorious "big-M"s to enforce logical constraints; the big-Ms led to weak optimality cuts to be added within the B&C procedure. We also proposed a B&B procedure which does not require big-Ms in the formulation. The B&B procedure begins by viewing the MMDP as $|\mathcal{M}|$ independent MDPs. The algorithm began by allowing each model of the MDP to have its own policy and sequentially added requirements that the decision rules for certain state-time pairs must agree in each model.

We presented the first numerical study of exact algorithms for MMDPs for realistic problem instances. To do so, we generated random MMDP instances of a machine repair problem to compare the computation time required to solve these problems using the extensive form, the B&C procedure, and the B&B procedure. Our computational experiments showed that the B&B solution method greatly outperforms the solution of the extensive form directly and with a B&C method. The B&B solution methods outperform both the extensive form and the B&C on all of our test cases. We conjecture that the big-M's in the formulation give rise to a poor LP relaxation which cause the MIP-based methods to not perform well. The B&B procedure was able to solve all but 5 of the 240 test instances to within 0.01% of optimality, while the other solution procedures were unable to solve any of the 240 instances to the same tolerance within the time limit. The worst-case optimality gap across all instances was 6.86% for the B&B method, 30.29% for the solution of the extensive form, and 34.33% for the B&C method. The average-case optimality gap across all instances was 0.06% for the B&B method, 12.22% for the solution of the extensive form, and 13.12% for the B&C method. In general, higher solution times for the B&B procedure resulted when there was higher variance among the models and when there are more models in the MMDP.

Our solution methods enabled us to investigate the impact of ambiguity on MDPs. For the machine maintenance instances, we considered the impact of the concentration of the transition probability models around their mean as well as the number of models used in the MMDP on the value of the MMDP approach in terms of value relative to the MVP and expected regret relative to each model's optimal policy. We found that the MMDP

approach was most beneficial when the DM has models that are quite different. When the models are similar, the MVP served as a good approximation, but the B&B procedure typically also solved the MMDP quickly in these cases.

Our study presented in this chapter has limitations. First, our B&C procedure does not include logic-based optimality cuts which could potentially improve its performance. However, our initial experiments with such cuts showed they do not provide significant enhancements to the B&C procedure. Second, in our computational experiments, we considered cases in which the rewards are the same in each model, however, there might be some situations in which a DM may want to consider model-specific rewards. Third, we reported the expected value of perfect information and the value of the MMDP for a particular instance of a machine repair problem; these values may depend on this problem's particular transition probability and reward structure and not be representative of all MDPs with ambiguity in the transition probabilities. Further, the transition probabilities in our experiments are sampled row-wise independently, but there may be value to investigating cases where ambiguity manifests itself through dependency across rows.

Our approach could be extended in several ways. First, we consider only finite-horizon MMDPs, but our decomposition algorithms might be extended to the infinite horizon setting. For instance, the B&B method might be extended to infinite-horizon MMDPs by using linear programming, value iteration, or policy iteration to solve the relaxations at each node in the B&B tree. It would be interesting to compare this approach to those proposed in Buchholz and Scheftelowitsch [14]. Second, we could further investigate other branching and node selection strategies to enhance the B&B procedure. Third, we do not assume any structure on the original MDP or its optimal policy. However, if we were to assume structure, such as monotonicity of the optimal policy, our algorithms might be able to be modified to exploit this structure for computational gain.

Chapter 4

Ambiguity-aware Multi-model Markov decision processes

4.1. Introduction

The main objective of this chapter is to introduce other ambiguity-aware formulations of the MMDP. As discussed in Chapter 2, the MMDP is a way to model ambiguity in MDPs through a finite number of models of the MDP ambiguous parameters. In Chapters 2 and 3, we presented the WVP wherein the DM seeks to maximize the expected rewards with respect to the different models:

$$\max_{\pi \in \Pi} \left\{ \mathbb{E}^{\xi} \left[v(\pi, \xi(m)) \right] \right\}, \tag{4.1}$$

where

$$v(\pi, \xi(m)) = \mathbb{E}^{\pi, P^m, R^m} \left[\sum_{t=1}^{T} r_t(s, \pi_t(s)) + r_{T+1}(s) \right], \tag{4.2}$$

and $\xi(m) = (P^m, R^m)$ represents a particular realization of the ambiguous transition probabilities and rewards for model $m \in \mathcal{M}$. Although, in general, the optimal policy for (4.1) may be history-dependent, as in the previous chapters we restrict our attention to the class of Markov deterministic (MD) policies.

The expectation in the objective (4.1) is appropriate for a DM who is risk-neutral to ambiguity. However, there is considerable evidence that some DMs may be risk-averse to the ambiguity which can affect decision-making [21]. For instance, Berger, Bleichrodt, and Eeckhoudt [6] showed that patients are less likely to opt for treatment when there

is ambiguity in terms of the effects of treatment. Therefore, it could be desirable to find a policy that offers more protection against the possible outcomes of the ambiguity and help guide DMs who exhibit some degree of ambiguity aversion. To do so, we can consider other ambiguity-aware formulations of the MMDP which reflect preferences beyond risk neutrality. Some alternative approaches to addressing risk include maximizing the worst-case (max-min), minimizing the maximum regret (min-max-regret), and percentile optimization (PercOpt). In the max-min approach, the DM seeks to find a policy that will maximize the expected rewards in the worst-case realization of the ambiguity, i.e.,

$$\max_{\pi \in \Pi} \min_{m \in \mathcal{M}} v(\pi, \xi(m)) \tag{4.3}$$

The min-max-regret criterion considers the performance relative to the best possible performance in that model. The regret for policy π in model m, $\ell(\pi, m)$ is the difference between the optimal value in model m and the value achieved by policy π in model m:

$$\ell(\pi, m) = \max_{\bar{\pi} \in \Pi} v(\bar{\pi}, \xi(m)) - v(\pi, \xi(m)) \tag{4.4}$$

The min-max-regret criterion then seeks to find the policy $\pi \in \Pi$ that minimizes the maximum regret:

$$\min_{\pi \in \Pi} \max_{m \in \mathcal{M}} \ell(\pi, m) \tag{4.5}$$

In the PercOpt approach, the DM selects a level of confidence $\epsilon \in [0, 1]$, and wants to maximize the ϵ-percentile of the value in the models, z:

$$\max_{z \in \mathbb{R}, \pi \in \Pi} z \text{ s.t. } \mathbb{P}\left(v(\pi, \xi(m)) \geq z\right) \geq 1 - \epsilon. \tag{4.6}$$

As before, we consider these preferences in the context of a finite-horizon, finite-state, finite-action, non-adaptive MMDPs. We show that the methods we described earlier in this thesis are easily modified to capture these other preferences towards ambiguity. However, the formulations described above are all NP-hard so we compare to two well-known formulations that can be solved in polynomial time. The first is the MVP-MMDP discussed in previous chapters, wherein all ambiguous parameters take on their mean values. The MVP-MMDP formulation can be solved using the standard backward recursion method [56]. The second is a formulation of the MMDP that satisfies the commonly employed (s, a)-rectangularity assumption. To ensure that the assumption is met, we project the MMDP's

MMDP Formulation	Objective Function	Description
MVP-MMDP	$\max_{\pi\in\Pi^{MD}} \left\{v(\pi, \mathbb{E}^{\xi}[\xi(m)])\right\}$	Maximizes the value in the MVP
WVP-MMDP	$\max_{\pi\in\Pi^{MD}} \left\{\mathbb{E}^{\pi,\xi}\left[v(\pi, \xi(m))\right]\right\}$	Maximizes the weighted value among models
Min-max-MMDP	$\max_{\pi\in\Pi^{MD}} \min_{m\in\mathcal{M}} v(\pi, \xi(m))$	Maximize the worst-case value among all models
Max-min-regret-MMDP	$\min_{\pi\in\Pi^{MD}} \max_{m\in\mathcal{M}} \ell(\pi, m)$	Minimize the maximum regret among all models
PercOpt-MMDP	$\max_{\pi\in\Pi^{MD}, z\in\mathbb{R}} z$ s.t. $\mathbb{P}\left(v(\pi, \xi(m)) \geq z\right) \geq 1-\epsilon.$	Maximize value z s.t. less that ϵ chance that $v(\pi, \xi(m)) < z$
(s,a)-rect-MMDP	$\max_{\pi\in\Pi^{MD}} \min_{P\in\mathscr{P}} \mathbb{E}^{P}\left[v(\pi)\right]$ $\mathscr{P} = \times_{s\in\mathcal{S}, a\in\mathcal{A}, t\in\mathcal{T}} \mathscr{P}_t(s,a)$	Maximize the worst-case in the (s,a)-rectangular finite scenario model

Table 4.1: A summary of the formulations of the MMDP considered.

parameters onto an (s,a)-rectangular ambiguity set. The (s,a)-rectangular MMDP (hereafter, (s,a)-rect-MMDP) was first proposed in Nilim and El Ghaoui [52] under the name of a *finite scenario model* and is easily solved using a backward recursion approach. We summarize the formulations considered in Table 4.1.

In this chapter, we present formulations of the MMDP that would be suitable to a DM whose ambiguity aversion can be represented by the models described above. The weighted value, max-min, min-max-regret, and PercOpt formulations of the MMDPs are all NP-hard [18, 40]. In Section 4.3 we summarize the existing MIP formulations for each objective function. We show that with modifications to our B&B solution method described in Chapter 3 we can exploit similar special structure of these new MMDP formulations for computational gain. In Section 4.4, we compare these ambiguity-aware formulations in two case studies. The first is a study of machine maintenance with ambiguity in the model of deterioration and the second is CVD management with ambiguity in how TC and HDL progress over time. We discuss the implications of our findings for decision-making under ambiguity in MDPs. We finish with a summary of our analysis and the most important findings in Section 4.5.

4.2. Background

Thus far, we have addressed ambiguity in an MDP by using the MMDP which allows for multiple models of the parameters to be incorporated into the optimization process. In Chapters 2 and 3, we have compared policies for the MMDP on the basis of the weighted value objective function. The WVP can be viewed as maximizing an expectation over the models in which the likelihood of a model corresponds to its weight (which may be viewed as maximizing the subjective expected utility [61]). The expectation in the objective function which is aligned with the view that the DM is risk-neutral to ambiguity in the MDP. However, some DMs might be risk-averse to ambiguity in the MDP.

Risk-aversion has been the standard paradigm for this early work on MDPs with ambiguity in the parameters. Much of the early work for MDPs has focused on determining the max-min policy when imprecise transition probabilities are allowed to vary within an ambiguity set [7]. Early work described the ambiguity sets using a polytope and sought to find policies that maximize the worst-case performance for MDPs whose parameters lie in these sets [60, 73], although these models tended to be computationally expensive. Givan, Leach, and Dean [25] described solution methods for the bounded-parameter MDP which is a special case of the earlier descriptions of MDPs with imprecise parameters. In the bounded-parameter MDP, the DM considers interval value functions representing the optimistic and pessimistic value functions for a given policy over the set of MDPs whose parameters lie within the bounds. Others extended this work to consider a multi-objective approach [62] when there are m sets of upper and lower bounds on the transition matrices. After that, more recent work in this area has taken on a robust optimization perspective wherein the ambiguous parameters of the MDP are allowed to vary within an ambiguity set. It has been shown that if the ambiguity set has the (s, a)-rectangularity property, these models are solved efficiently using a modification to standard solution methods [35, 36, 52]. However, there has been some concern that the resulting policies are overly conservative, i.e., the formulation focuses on the worst possible outcome wherein all of the random variables take on their worst-case values simultaneously which is an infrequent event in most practical situations; thus, a stream of literature has followed that with the aim of limiting the conservativeness of the policy. One area of research has been to construct tractable variations of the rectangular ambiguity set in a max-min framework [30, 45, 74, 80]. Others have modified formulation to consider regret or otherwise alter the objective to obtain a policy that accounts for ambiguity without being overly conservative [1, 24, 79]. Others

Objective for MMDP	**Complexity** ($\Pi = \Pi^{MD}$)	**Reduction**
WVP-MMDP	NP-hard	3-CNF-SAT [40, 70]
Max-min-MMDP	NP-hard	3-CNF-SAT [40]
Min-max-regret-MMDP	NP-hard	3-CNF-SAT [40]
PercOpt-MMDP	NP-hard	3-SAT [18]
(s, a)-rect MMDP	Polynomial	[52]

Table 4.2: A summary of the complexity results related to ambiguity-aware MMDPs

still have handled ambiguous parameters by assuming some distributional information on the ambiguous parameters and modifying the objective function in the max-min function [18, 41, 78].

In contrast to the work described above which considers a continuous set of transition kernels, this thesis has treated ambiguity in MDP by creating multiple discrete plausible models of the MDP. We term these *models*, although can be thought of as analogous to finite scenarios in stochastic programming. The interpretation of the models as scenarios promotes the view of the weights in the MMDP as the likelihood of the corresponding model and the WVP as the problem of maximizing an expectation over the set of models. In this chapter, we extend the formulation of the MMDP to consider other preferences towards ambiguity in the MDP's parameters.

We now summarize the most closely related work in the literature, especially other recent formulations of MDPs that incorporate parameter ambiguity through a finite set of models. A finite scenario model was considered in Nilim and El Ghaoui [52] under the assumption that the models would satisfy the $(s, a)-$rectangularity property which is computationally attractive. Le Tallec [40] considered finite scenarios in an MMDP without the (s, a)-rectangularity assumption but their analysis focused on the complexity of MDPs problems with ambiguous transition probabilities. The authors showed that the max-min, min-max-regret, and weighted value cases for non-adaptive MMDPs are all NP-complete in general. Delage and Mannor [18] considered a finite number of models in their proof that percentile optimization is NP-hard in general and then focused on methods for solving the model under row-wise independence assumptions. Despite these eariler works, efforts to solve these problems have largely been left untouched until only recently. Ahmed et al. [1] described the min-max-regret-MMDP. They show that, in general, the optimal policy for this problem may be randomized, but also present a MIP formulation that can be

Objective for MMDP	Finite Horizon	Infinite Horizon
WVP-MMDP	MIP [Ch. 2], B&C, B&B [Ch. 3]	MIP [14]
Max-min-MMDP	MIP, B&B [Ch. 4]	
Min-max-regret-MMDP	MIP [1], B&B [Ch. 4]	
PercOpt-MMDP	MIP, B&C [49], B&B [Ch. 4]	
(s, a)-rect MMDP	Modified Backwards Recursion [52]	Modified Bellman Recursion [52]

Table 4.3: A summary of the solution methods used to solve ambiguity-aware MMDPs. If there is no known solution method for this case, we leave the cell empty.

used to solve the problem for the class of Markov deterministic policies. We propose a modification to the B&B algorithm presented in Chapter 3 which could be used to solve this formulation of the MMDP. We also consider a PercOpt approach for a finite number of models. Concurrently with Merakli [49], we considered this problem in the setting where the models are treated random variables with finite support and probabilities corresponding to their weights. In this setting, we seek to maximize the ϵ-quantile. Merakli [49] provide a MIP formulation as well as a corresponding B&C scheme similar to the one proposed for the WVP in Chapter 3. For the PercOpt MMDP, also provide a modification to the B&B solution method. A summary of the complexity results for these problems is given in 4.2, and a summary of the existing solution methods for solving these MMDPs is given in Table 4.3.

As we presented in Chapter 2, the WVP of a non-adaptive MMDP can be viewed as a two-stage stochastic integer program. In the classical two-stage stochastic program, the second-stage rewards depend on the scenario and are uncertain at the time the DM selects first-stage decisions. The classical objective function is to maximize an objective function which a first-stage reward function component and an expected second-stage reward component [9]. The expectation serves as a risk-mapping from the random variable representing the second-stage rewards to a scalar which can be used to compare decisions. The expectation risk-mapping reflects a risk-neutral approach to ambiguity, but other risk-mappings have been proposed as a way to reflect other preferences for the ambiguity that arises due to the lack of knowledge around model parameters. Others have proposed alternative risk mappings in the two-stage stochastic integer programming setting and provided corresponding decomposition methods for solving them which could presumably be used to solve the MIP formulations of the MMDPs with alternative risk preferences [19, 43]. How-

ever, we consider the study of possible decomposition methods for the MIP formulations to be beyond the scope of this chapter.

In this chapter, we make the following novel contributions. We summarize the existing formulations of ambiguity-aware MMDPs and their computational complexity. We show that these formulations can all be solved in a common algorithmic framework which leverages the special decomposable nature of MMDPs discussed in Chapter 3. We provide a computational study of these formulations for finite-horizon MMDPs for case studies involving medical decision making and machine repair, similar to the examples of Chapters 2 and 3. Finally, we compare the optimal solutions to the ambiguity-aware MMDPs formulations with a formulation that satisfies the commonly employed (s, a)-rectangularity property.

4.3. Solution methods

In this section, we describe solution methods that can be used to solve the ambiguity-aware formulations of the MMDPs. For those formulations with existing MIP formulations, we still provide the formulation here in our notation for completeness. These formulations can easily be provided to a standard commercial solver; however, the presence of binary decision variables and big-Ms is likely to limit the solutions via commercial solvers to very small instances, as shown for the WVP-MMDP in Chapter 3. Therefore, for each formulation, we also provide a modification to the B&B algorithm presented in Chapter 3 which leverages the separable structure of MMDPs.

4.3.1. Max-min-MMDP

The max-min MMDP as presented in (4.3) can be formulated as the following MIP:

$$\max_{\pi,\, v,\, w} \quad w \tag{4.7a}$$

$$\text{s.t.} \quad \sum_{a\in\mathcal{A}} \pi_t(a|s) = 1, \qquad \forall s \in \mathcal{S}, t \in \mathcal{T}, \tag{4.7b}$$

$$M\pi_t(a|s) + v_t^m(s) - \sum_{s'\in\mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \le r_t^m(s,a) + M, \tag{4.7c}$$

$$\forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T},$$

$$v_{T+1}^m(s) \le r_{T+1}^m(s), \forall m \in \mathcal{M}, s \in \mathcal{S}, \tag{4.7d}$$

$$w - \sum_{s \in \mathcal{S}} \mu_1^m(s) v_1^m(s) \leq 0, \qquad \forall m \in \mathcal{M}, \tag{4.7e}$$

$$\pi_t(a|s) \in \{0,1\}, \quad \forall a \in \mathcal{A}, s \in \mathcal{S}, t \in \mathcal{T}, \tag{4.7f}$$

$$v_t^m(s) \in \mathbb{R}, \qquad \forall m \in \mathcal{M}, s \in \mathcal{S}, t \in \mathcal{T}, \tag{4.7g}$$

$$w \in \mathbb{R}. \tag{4.7h}$$

The formulation in (4.7) represents the policy through the binary variables π and the value functions for each model, state, and decision epoch through the continuous variables v. The continuous variable w represents the worst-case model value and takes on the appropriate value due to constraint (4.7e).

However, rather than solving the MIP directly, one can solve the problem using a modification of the B&B algorithm described in Chapter 3. To do so, one needs to modify the bounding approach to reflect the best possible worst-case model value that comes from completing a partial policy. The upper bound on the worst-case value is determined by selecting the value for the model with the smallest upper bound found by solving the relaxation. That is, Step 14 in Algorithm 5 (presented in Chapter 3), should be modified to

$$\bar{z}^i \leftarrow \min_{m \in \mathcal{M}} v^m(\hat{\pi}^i) \tag{4.8}$$

to reflect the appropriate upper bound.

4.3.2. Min-max-regret-MMDP

We now present a MIP formulation to solve the min-max-regret MMDP as in (4.5). Ahmed et al. propose an equivalent formulation to the one below. We present the MIP formulation using the notation of this thesis for convenience:

$$\min_{\pi, v, w, \bar{v}} \quad w \tag{4.9a}$$

$$\text{s.t.} \quad \sum_{a \in \mathcal{A}} \pi_t(a|s) = 1, \qquad \forall s \in \mathcal{S}, t \in \mathcal{T}, \tag{4.9b}$$

$$M\pi_t(a|s) + v_t^m(s) - \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \leq r_t^m(s,a) + M, \tag{4.9c}$$

$$\forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T},$$

$$v_{T+1}^m(s) \leq r_{T+1}^m(s), \forall m \in \mathcal{M}, s \in \mathcal{S}, \tag{4.9d}$$

$$\bar{v}_t^m(s) - \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) \bar{v}_{t+1}^m(s') \geq r_t^m(s,a), \tag{4.9e}$$

$$\forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T},$$

$$\bar{v}_{T+1}^m(s) \geq r_{T+1}^m(s), \forall m \in \mathcal{M}, s \in \mathcal{S}, \tag{4.9f}$$

$$\sum_{s \in \mathcal{S}} \mu_1^m(s) \bar{v}_1^m(s) - \sum_{s \in \mathcal{S}} \mu_1^m(s) v_1^m(s) - w \leq 0, \qquad \forall m \in \mathcal{M}, \tag{4.9g}$$

$$\pi_t(a|s) \in \{0,1\}, \quad \forall a \in \mathcal{A}, s \in \mathcal{S}, t \in \mathcal{T}, \tag{4.9h}$$

$$v_t^m(s), \bar{v}_t^m(s) \in \mathbb{R}, \qquad \forall m \in \mathcal{M}, s \in \mathcal{S}, t \in \mathcal{T}, \tag{4.9i}$$

$$w \in \mathbb{R}. \tag{4.9j}$$

In the MIP formulation above, we define $\pi \in \{0,1\}^{|\mathcal{S}| \times |\mathcal{A}| \times |\mathcal{T}|}$ and $v \in \mathbb{R}^{|\mathcal{M}| \times |\mathcal{S}| \times |\mathcal{A}| \times |\mathcal{T}|}$ as above representing the policy and value function variables of the MMDP, respectively. In the min-max-regret formulation, we introduce continuous variables $\bar{v}$ to represent the optimal value functions for each model. Notices that these continuous variables are independent of the policy variables π. The constraints in (4.9e) and (4.9f) ensure that the variables $\bar{v}^m$ take on the optimal value functions corresponding to model $m \in \mathcal{M}$. We also introduce a continuous variable w to represent the worst-case regret among all of the models. Constraints (4.9g) serve as epigraph constraints ensuring that w will take on the value of the largest regret among all of the models. To minimize the value of w, the variables v^m will attempt to take on values as close to their $\bar{v}^m$ counterparts, so long as the constraints in (4.9c) and (4.9d), which enforce that these value functions correspond to policy π, are satisfied.

Rather than solving the MIP directly, the B&B algorithm described in Chapter 3 can be modified to apply to this problem. To do so, the B&B requires a modification to

reflect that this is a minimization problem. Further, one needs to modify the bounding approach to reflect the best possible worst-case model regret that comes from completing a partial policy. The bound is easily found after solving the relaxation, which is done in the same manner as before. However, at a given node, the lower bound needs to reflect the worst-case regret by determining which model's value in the relaxation is furthest from the corresponding model's optimal value. To reflect this modification to the lower bound, Step 14 in Algorithm 5 (presented in Chapter 3), should be modified to

$$\bar{z}^i \leftarrow \min_{m \in \mathcal{M}} \left(\max_{\bar{\pi} \in \Pi} \{v(\bar{\pi}, \xi(m))\} - v^m(\hat{\pi}^i) \right) \tag{4.10}$$

Then, one can remove nodes from consideration in the B&B procedure by pruning any node whose lower bound is greater than the incumbent solution in the minimization problem.

4.3.3. PercOpt-MMDP

In the following MIP formulation, we define π and v as above representing the policy and value function variables, respectively. In the PercOpt formulation, we introduce binary variables y_m for each model which take on a value of 1 if the value in model m is at least z and zero otherwise. Constraint (4.11e) ensures that all models with values greater than or equal to z will have corresponding y_m variables equal to 1 and all models with a value less than z will have a y_m value of 0. Constraint (4.11f) ensures that there is no more than ϵ chance that the models have values less than z.

$$\max_{\pi, v, y, z} \quad z \tag{4.11a}$$

$$\text{s.t.} \quad \sum_{a \in \mathcal{A}} \pi_t(a|s) = 1, \qquad \forall s \in \mathcal{S}, t \in \mathcal{T}, \tag{4.11b}$$

$$M\pi_t(a|s) + v_t^m(s) - \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \leq r_t^m(s,a) + M, \tag{4.11c}$$

$$\forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T},$$

$$v_{T+1}^m(s) \leq r_{T+1}^m(s), \forall m \in \mathcal{M}, s \in \mathcal{S}, \tag{4.11d}$$

$$z + My_m - \sum_{s \in \mathcal{S}} \mu_1^m(s) v_1^m(s) \leq M, \qquad \forall m \in \mathcal{M}, s \in \mathcal{S}, \tag{4.11e}$$

$$\sum_{m\in\mathcal{M}} \lambda_m y_m \geq 1-\epsilon \quad \forall m \in \mathcal{M}, \tag{4.11f}$$

$$\pi_t(a|s) \in \{0,1\}, \quad \forall a \in \mathcal{A}, s \in \mathcal{S}, t \in \mathcal{T}, \tag{4.11g}$$

$$y_m \in \{0,1\}, \quad \forall m \in \mathcal{M}, \tag{4.11h}$$

$$v_t^m(s) \in \mathbb{R}, \quad \forall m \in \mathcal{M}, s \in \mathcal{S}, t \in \mathcal{T}, \tag{4.11i}$$

$$z \in \mathbb{R}. \tag{4.11j}$$

Similar to the cases above, rather than solving the MIP directly, one can solve the problem using a modification of the B&B algorithm described in Chapter 3. To do so, the upper bound on the ϵ-percentile is determined by selecting the value for the model with the smallest upper bound found by solving the relaxation. Step 14 in Algorithm 5 (presented in Chapter 3) should be modified to reflect the appropriate upper bound. The upper bound $\bar{z}^i$ is easily found by sorting the models in increasing order by the value functions $v^m(\hat{\pi}^i)$ to obtain an order statistic $m_{(1)}, \ldots, m_{(|\mathcal{M}|)}$. Then, one can select $\bar{z}^i$ to be $v^{m_{(x)}}(\hat{i})$ where x is the smallest value such that $\sum_{m=m_{(1)}}^{m_{(x-1)}} \lambda_m \leq \epsilon$.

4.3.4. (s,a)-rect-MMDP

We compare the solution of the formulations of the MMDP to the (s,a)-rectangular finite scenario model described in Nilim and El Ghaoui [52]. This model imposes the rectangularity assumption which has been popular in past work, in part because it imposes independence between rows of the transition probability matrix, making the resulting problem solvable in polynomial time. To construct the (s,a)-rect-MMDP, we project the parameters in the MMDP onto an (s,a)-rectangular ambiguity set. The projection is done by constructing a ambiguity set that is independently constructed for each (s,t,a)-tuple for $(s,t,a) \in \mathcal{S} \times \mathcal{T} \times \mathcal{A}$:

$$\mathscr{P} = \times_{s\in\mathcal{S}, a\in\mathcal{A}, t\in\mathcal{T}} \mathscr{P}_t(s,a)$$

and

$$\mathscr{R} = \times_{s\in\mathcal{S}, a\in\mathcal{A}, t\in\mathcal{T}} \mathscr{R}_t(s,a)$$

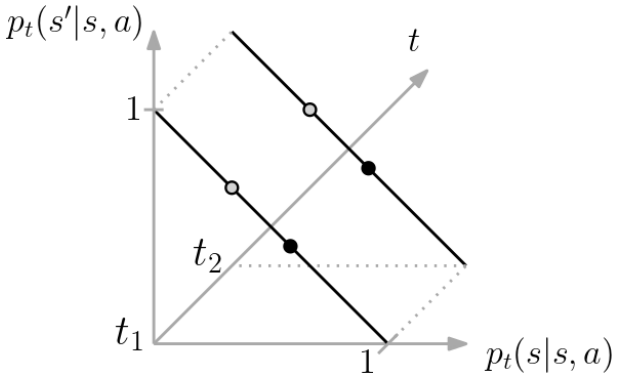


Figure 4.1: An illustration of how an MMDP would be projected on an (s,a)-rectangular ambiguity set. Suppose that in the MMDP, the DM knows that the transition probability row for (s,a) is the same at time t_1 and t_2, but is unsure of its precise value. A MMDP representation of this ambiguity would be to specify two possible values: $p_t^1(\cdot|s,a)$ and $p_t^2(\cdot|s,a)$. This is illustrated in the figure: the grey dots would represent $p_t^1(\cdot|s,a)$ and the black dots would represent $p_t^2(\cdot|s,a)$. Under the (s,a)-rectangular ambiguity set model, there is an adversary that is able able to pick either $p_{t_1}^1(\cdot|s,a)$ or $p_{t_1}^2(\cdot|s,a)$ at time t_1 and either $p_{t_2}^1(\cdot|s,a)$ and $p_{t_1}^2(\cdot|s,a)$ at time t_2.

with

$$\mathscr{R}_t(s,a) = \left\{r_t^1(s,a), r_t^2(s,a), \ldots, r_t^{|\mathcal{M}|}(s,a)\right\}, \forall s \in \mathcal{S}, t \in \mathcal{T}, a \in \mathcal{A},$$

and

$$\mathscr{P}_t(s,a) = \left\{p_t^1(\cdot|s,a), p_t^2(\cdot|s,a), \ldots, p_t^{|\mathcal{M}|}(\cdot|s,a)\right\}, \forall s \in \mathcal{S}, t \in \mathcal{T}, a \in \mathcal{A}.$$

The resulting ambiguity set is discrete and (s,a)-rectangular. The goal of the DM is then to solve the robust MDP formulation:

$$\max_{\pi \in \Pi} \min_{P \in \mathscr{P}, R \in \mathscr{R}} \mathbb{E}^{\pi,P,R}\left[\sum_{t=1}^{T} r_t(s,\pi_t(s)) + r_{T+1}(s)\right], \tag{4.12}$$

Figure 4.1 illustrates the projection of MMDP parameters onto a (s,a)-rectangular ambiguity set. Algorithm 6 describes how to solve (4.12) in polynomial time.

4.4. Case studies

In this section, we provide two case studies to analyze the impact of ambiguity on decision-making and how a DM's preference may influence the best course of action. The first is an MMDP used to optimize repairs in a machine maintenance setting where the DM makes

Algorithm 6: Modified backwards recursion as in Nilim and El Ghaoui [52] to solve the (s, a)-rect-MMDP

Data: MMDP
Result: The policy $\pi^{WC} = (\pi_1^{WC}, \ldots, \pi_T^{WC}) \in \Pi^{MD}$

Let $v_{T+1}^{WC}(s_{T+1}) = \min_{m \in \mathcal{M}}\{r_{T+1}^m(s_{T+1})\}$
$t \leftarrow T$
while $t \geq 1$ **do**
  for *Every state* $s_t \in \mathcal{S}$ **do**

$$\pi_t^{WC}(s_t) \leftarrow \underset{a_t \in \mathcal{A}}{arg\,max} \left\{ \min_{m \in \mathcal{M}} \left(r_t^m(s_t, a_t) + \sum_{s_{t+1} \in \mathcal{S}} p_t^m(s_{t+1}|s_t, a_t) \hat{v}_{t+1}^{WC}(s_{t+1}) \right) \right\} \tag{4.13}$$

$$v_t^{WC}(s_t) = \max_{a_t \in \mathcal{A}} \left\{ \min_{m \in \mathcal{M}} \left(r_t^m(s_t, a_t) + \sum_{s_{t+1} \in \mathcal{S}} p_t^m(s_{t+1}|s_t, a_t) \hat{v}_{t+1}^{WC}(s_{t+1}) \right) \right\} \tag{4.14}$$

  end
  $t \leftarrow t - 1$
end

many recurring decisions over time. The second considers the optimal timing of statin therapy for the prevention of CVD. In this setting, the initiation and intensification of statins are considered irreversible decisions, as is consistent with the clinical recommendations for those medications [68].

In each case study, we use the alternative formulations of the MMDP described in Section 4.1 to address ambiguity that arises in the transition dynamics of the MDP. We solve each formulation using the modified B&B algorithms described in Section 4.3, or their backward recursion algorithm when appropriate. Then, we compare the resulting policies in terms of their performance with respect to each of the models, as well as their performance with respect to the other risk-preferences towards ambiguity.

4.4.1. Case study: Machine maintenance

In this section, we consider the machine maintenance problem that we described in Chapter 3 to compare the ambiguity-aware formulations of the MMDP. For the ease of reading, we briefly restate the problem below, and we describe our findings afterward.

MMDP formulation

We consider the machine maintenance problem with 6 states, 3 actions, and a uniform initial state distribution. States 0 to 5 represent the quality of the machine, with 0 being the highest quality and 5 being the lowest. There are 3 actions: *Do Nothing*, *Repair Option 1*, and *Repair Option 2*. Repair Option 1 provides lower improvement at lower cost and Repair Option 2 provides higher improvement at the higher cost. Figure 3.1 illustrates the expected values of the transition probabilities for actions *Do Nothing* (Action 0) and *Repair Option 1* (Action 1). The transition probabilities for *Repair Option 2* are similar to *Repair Option 1* but allow for the machine to potentially improve two states in between epochs. Performing *Repair Options 1* and *2* incur cost penalties of 5 and 8, respectively. We assume that the cost to operate the machine is dependent on the quality of the machine and that when the machine reaches the age of 7, it will be replaced completely. The planning horizon consists of 6 decision epochs, and we ignore the cost to replace the machine at time 7.

Model generation We construct an MMDP by sampling transition matrices from a Dirichlet distribution. The Dirichlet distribution used has parameters $(\alpha\bar{p}_1, \ldots, \alpha\bar{p}_{|\mathcal{S}|})$ for each row of the transition probability matrix where the base measure of the distribution, $(\bar{p}_1, \ldots, \bar{p}_{|\mathcal{S}|})$, corresponds to the mean values of the parameters of that row. We use the same mean value parameters as in Chapter 3. In this chapter, we consider three MMDP instances of this problem, each with $|\mathcal{M}| = 100$. The first is an instance where the transition probabilities are sampled from a Dirichlet distribution with a concentration parameter of $\alpha = 10$. In this MMDP, the parameters of each model are not closely concentrated around the parameters describing the MVP. The second and third are instances in which the concentration parameter of the Dirichlet distribution is $\alpha = 20$ and 100, respectively. For these instances, the models' parameters are more closely concentrated around the MVP's parameters. These distributions are illustrated in Figure 4.2. We have selected these instances to illustrate how ambiguity and the DM's preferences towards this ambiguity may or may not influence the best course of action. Figure 4.2 illustrates these distributions.

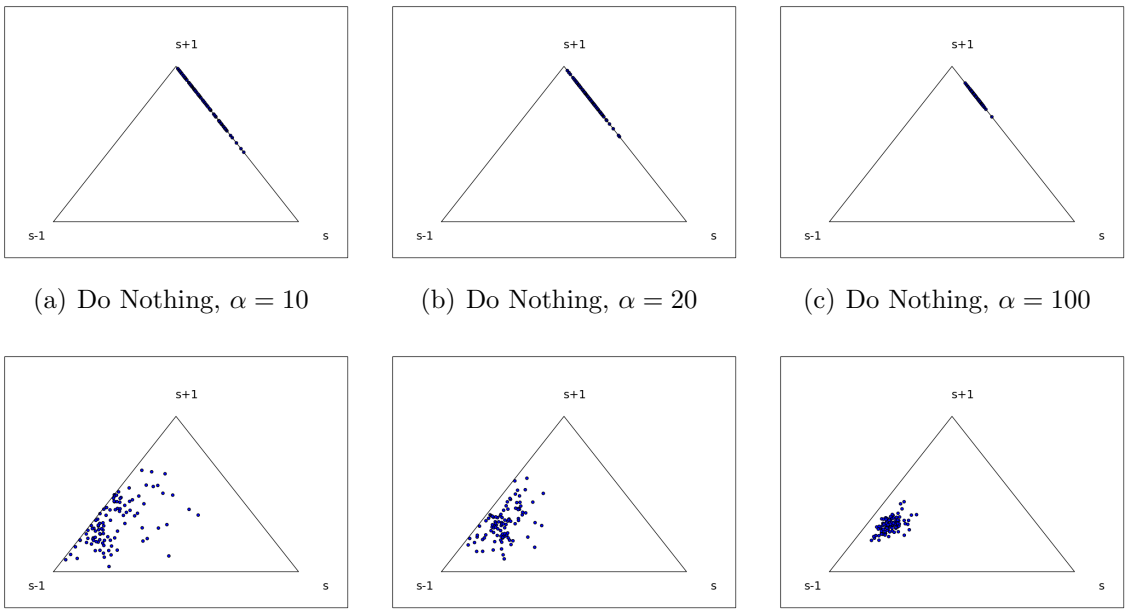
Experiments We solved the MMDPs described above using the modified B&B algorithms and the (s, a)-rectangular projection using Algorithm 6 for the 3 values of the concentration parameter. For each MMDP formulation, we solved to within an optimality gap of 2%. We solved all instances and reported their gaps after 300 seconds. All instances were solved

on a Macintosh MacBook Pro with a 2.7 GHz CPUs and 16 GB of memory using the B&B algorithm implemented in C++ using Xcode Version 10.0. For the problem instance where $\alpha = 100$, all MMDP formulations were solved in less than 0.01 CPU seconds, except the min-max-regret-MMDP which did not find a better incumbent solution and finished with a gap of 5.7%. For the problem instance with $\alpha = 20$, the MVP-,(s, a)-rect, WVP-MMDP were all solved in under 3 seconds. The max-min-MMDP took 15.0 seconds and the PercOpt-MMDP and min-max-regret-MMDP were unable to be solved in the time limit, finishing with gaps of 2.2% and 40.4%, respectively. For $\alpha = 10$, the MMDP formulations took the most time to solve. After 300 seconds, the following formulations had these respective gaps: WVP-MMDP: 2.4%, max-min-MMDP: 10.6%, min-max-regret-MMDP: 50.7%, and PercOpt: 6.2%. It appears that higher variance in the MMDP's parameters leads to weaker upper bounds and degrades the performance of the B&B algorithms. In the results to follow we present the results for instances that were eventually solved while noting there is some bias to presenting results for this subset of "solvable" model instances.

Results

Policy comparisons We now describe the results of our experiments. First, we compare each of the policies obtained solving the alternative formulation of the MMDP and evaluating those policies in each model of the MMDP.

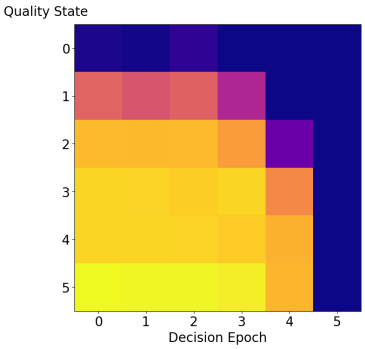
Figures 4.3, 4.4, and 4.5 illustrate the policies recommended for each of the formulations of the MMDP for different values for concentration parameters values $\alpha = 10,\ 20$, and 100, respectively. First, consider Figure 4.3 which illustrates the policies for the instance when there is the highest amount of dispersion of the models' parameters ($\alpha = 10$). Figure 4.3(a) illustrates a "heatmap" of the wait-and-see policies. The darkest parts of the figure illustrate states and decision epochs for which all of the models agree that the machine does not undergo a repair. The lightest portion represents the states and decision epochs for which all models agree that the machine should undergo the most extreme repair. The gradient of colors in between represents the amount of disagreement among the wait-and-see policies with lighter colors suggesting more of the policies recommend a repair in this state and decision epoch while darker colors suggest more of the policies recommend the machine forgoes a repair at the state and decision epoch. Figures 4.3(b)-4.3(e) illustrate the MMDP policies. The MVP- and WVP-MMDPs provide the same solution, so for this instance the VSS=0. We observe that these policies recommend major repairs for machines



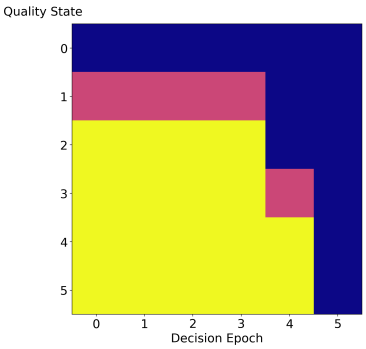
(a) Do Nothing, $\alpha = 10$ (b) Do Nothing, $\alpha = 20$ (c) Do Nothing, $\alpha = 100$

(d) Repair, $\alpha = 10$ (e) Repair, $\alpha = 20$ (f) Repair, $\alpha = 100$

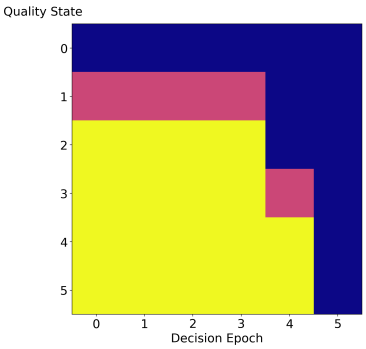
Figure 4.2: Samples from the Dirichlet distributions used in the machine repair case study. Each figure represents 100 samples from the Dirichlet distributions for the corresponding action and concentration parameter, α, that are used to generate the models for the case study. The triangle represents the probability simplex defining the transition probabilities from state s to $s-1$ (lower left corner of the triangle), s (lower right corner) and $s+1$ (top corner). The top row 4.2(a)-4.2(c) corresponds to transitions for the action to "Do Nothing"; these probabilities were drawn from a Dirichlet distribution with a mean $(0, 0.2, 0.8)$. These figures represents samples from the Dirichlet distribution with parameters $(0, 0.2\alpha, 0.8\alpha)$ for $\alpha = 10,\ 20$ and 100, respectively. The bottom row 4.2(d)-4.2(f) corresponds to transitions for the action to "Repair" with option 1; these probabilities were drawn from a Dirichlet distribution with a mean $(0.6, 0.1, 0.3)$. Each figure represents samples from the corresponding Dirichlet distribution with parameters $(0.6\alpha, 0.1\alpha, 0.3\alpha)$ for $\alpha = 10,\ 20$ and 100, respectively.



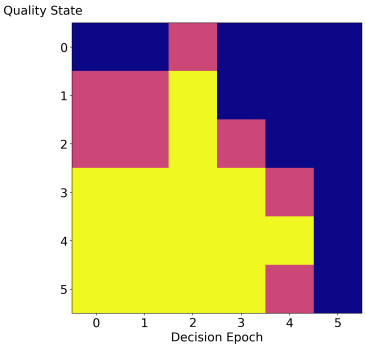
(a) Heatmap of wait-and-see



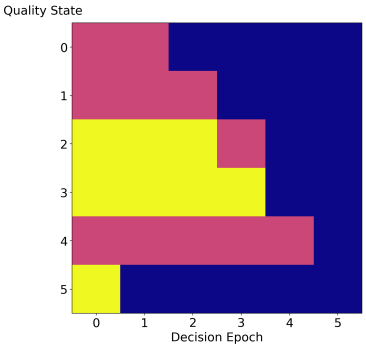
(b) The MVP-MMDP policy



(c) The WVP-MMDP policy

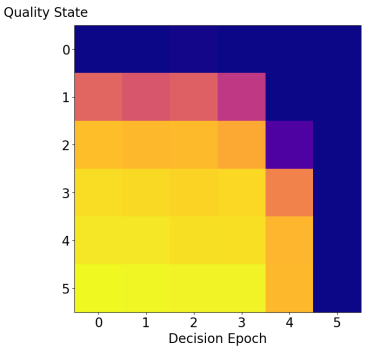


(d) The max-min-MMDP policy

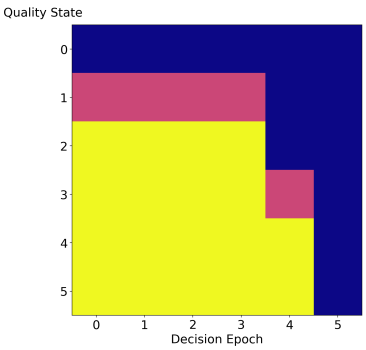


(e) The (s, a)-rect-MMDP policy

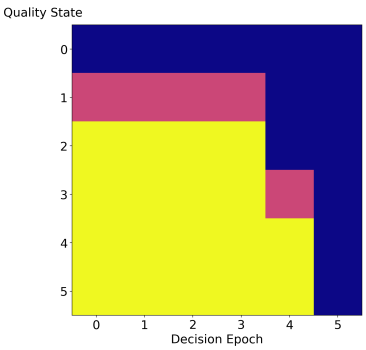
Figure 4.3: An illustration of the ambiguity-aware MMDP policies for the various formulations of the machine maintenance MMDP when the models are generated from a Dirichlet distribution with a concentration parameter of $\alpha = 10$. The lightest color corresponds to a major repair, the darkest action corresponds to doing nothing.



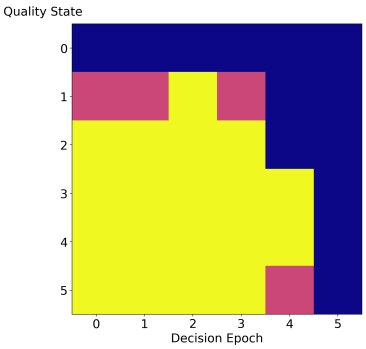
(a) Heatmap of wait-and-see



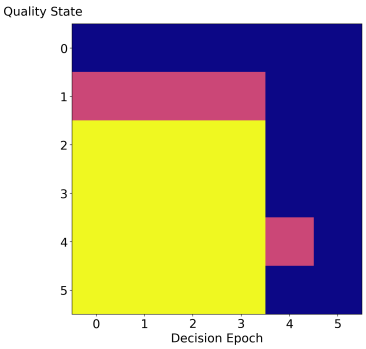
(b) The MVP-MMDP policy



(c) The WVP-MMDP policy

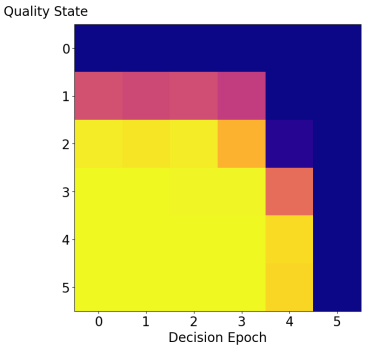


(d) The max-min-MMDP policy

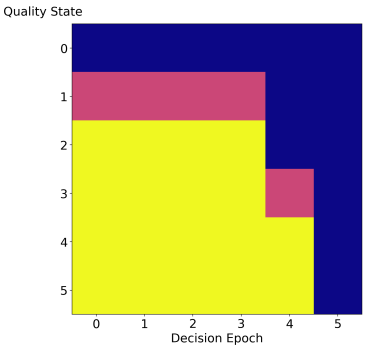


(e) The (s, a)-rect-MMDP policy

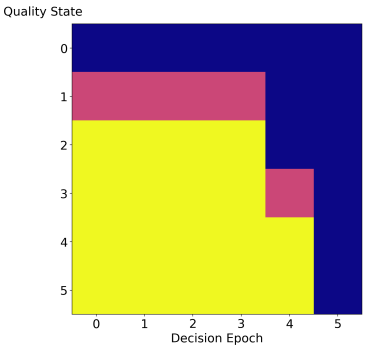
Figure 4.4: An illustration of the ambiguity-aware MMDP policies for the various formulations of the machine maintenance MMDP when the models are generated from a Dirichlet distribution with a concentration parameter of $\alpha = 20$. The lightest color corresponds to a major repair, the darkest action corresponds to doing nothing.



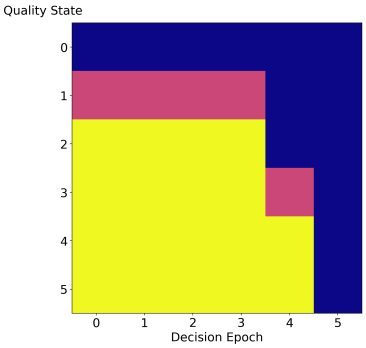
(a) Heatmap of wait-and-see



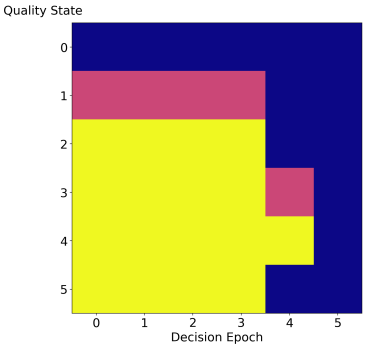
(b) The MVP-MMDP policy



(c) The WVP-MMDP policy



(d) The max-min-MMDP policy



(e) The (s, a)-rect-MMDP policy

Figure 4.5: An illustration of the ambiguity-aware MMDP policies for the various formulations of the machine maintenance MMDP when the models are generated from a Dirichlet distribution with a concentration parameter of $\alpha = 100$. The lightest color corresponds to a major repair, the darkest action corresponds to doing nothing.

at the beginning of the horizon. However, older machines and machines that are in better quality states should forgo repairs. The max-min-MMDP and the (s, a)-rect-MMDP are presented in Figure 4.3(d) and 4.3(e), respectively. The max-min-MMDP has a similar trend of recommending newer machines in worse quality states for repair and forgoing maintenance for the best quality machines and those that are older. However, this policy is more aggressive in terms of repairs in the middle of the horizon. The (s, a)-rect-MMDP has similar trends as before for machines that are in good quality. However, for the worse quality states, the (s, a)-rect-MMDP is less aggressive in terms of repairs. One possible explanation for this is that the DM is accounting for a case where the deterioration of machines in the worst quality states is so rapid that the repairs are not worth the cost.

Now consider Figure 4.4 which considers a case with less dispersion. We observe that the MVP and WVP agree again. In this instance, the max-min-MMDP agrees on which states to forgo repairs but differs in some aspects as far at what machine quality states should undergo major repairs and when. The (s, a)-rect-MMDP agrees with the MVP solution in terms of a maintenance plan in the beginning of the planning horizon, but tends to forgo repairs for older machines.

Finally, consider Figure 4.5 which illustrates the case with the least dispersion of the models' parameters. In Figure 4.5(a), we observe that there is less disagreement among the wait-and-see policies. Most models agree that repairs should be done for machines that are in poor quality at the beginning of the planning horizon, but there is still some disagreement as far as the repairs for states 1 and at decision epoch 4. We observe that again the MVP and WVP are unchanged from the MVP and WVP for $\alpha = 10$ and $\alpha = 20$. In this instance, these policies are the same as the max-min-MMDP model policy. The (s, a)-rect-MMDP policy differs only it its recommendation for quality state 5 in decision epoch 4.

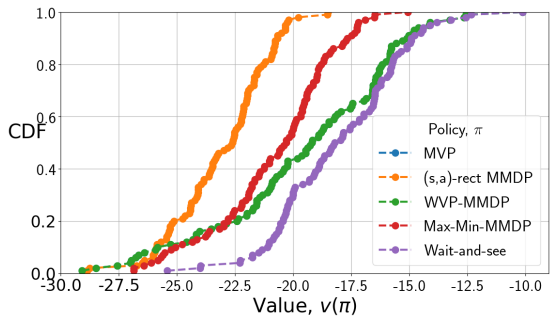
In summary, a higher variance in the models' transition probabilities results in more variation in the recommendations resulting from these models. However, we find that the WVP-MMDP and the MVP-MMDP tend to recommend the same course of action. We will illustrate the performance of these policies in the following section.

Value function comparisons Figures 4.6 shows the distributions of the value function based on the 100 models used in the various MDP formulations for the different levels of dispersion (as measured by the concentration parameter α) and the different MMDP policies. The distributions of the value function are illustrated via their cumulative dis-

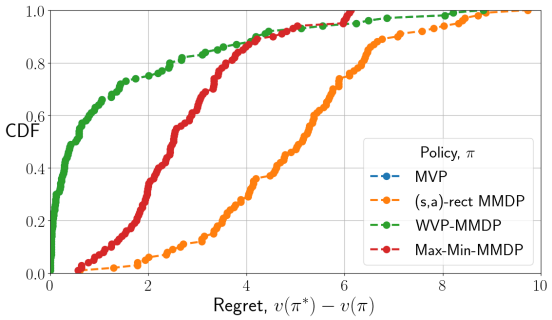
tribution function (CDF) where each model is considered to be equally likely. We also present the results in terms of the distribution of regret as illustrated by the CDF.

First, consider Figure 4.6(a) which shows the CDF of the value function in the MMDP with a high dispersion among the models parameters ($\alpha = 10$). The "wait-and-see" line represents the distribution of the optimal value functions for each model and serves as an upper bound on the achievable value function distribution. The max-min-MMDP policy does fare better in terms of the worst-case model with a cost of 26.9 and does perform better than the WVP-policy in terms of the 18th percentile. However, we observe that this policy underperforms with respect to the WVP in terms of the rest of the value function distribution. We also see the that WVP fares only slightly worse than the (s, a)-rect-MMDP in terms of worst-case model performance (a cost of 29.1 for WVP compared to 28.9 for (s, a)-rect-MMDP). The (s, a)-rect-MMDP does improve performance for the parts of the value function distribution with higher costs. However, the WVP performs much better in terms of other parts of the distribution. Figure 4.6(c) and Figure 4.6(e) show these distributions for the higher values of the concentration parameter. We observe that for higher values of the concentration parameter (i.e., when the parameters are more closely concentrated around their mean values), the (s, a)-rect-MMDP is a better approximation of the worst-case value function. However, the MVP-MMDP also performs well in these cases and achieves a worst-case model value that is comparable to the (s, a)-rect-MMDP.

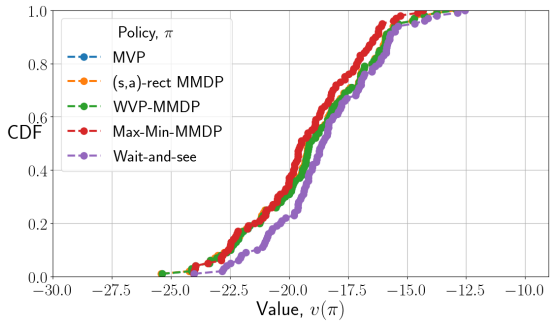
Figure 4.6(b) shows the distribution of the value functions in terms of the regret in the 100 models for each policy, as illustrated by the CDF. A vertical line at Regret = 0 would represent that the policy achieved zero regret in all models. We observe that the max-min-MMDP does better in terms of worst-case regret than the WVP (6.1 for max-min-MMDP vs. 8.8 for WVP-MMDP). The (s, a)-rect-MMDP fares the far worse in terms of regret, underperforming both the max-min-MMDP and the WVP-MMDP. The (s, a)-rect-MMDP has a maximum regret of 9.7 which is worse than either of the other MMDP policies. Thus, the effectiveness of the (s, a)-rect-MMDP in managing the impact of ambiguity may be on par with the effectiveness of the MVP. When the model parameters are more closely concentrated around their mean values, the policies tend to perform similarly. In Figure 4.6(d) illustrates the distributions for $\alpha = 20$. We observe that for the max-min-MMDP does achieve the best worst-case value, but this comes at a cost of sacrificing performance for many other models and in this case, the policy actually does worse on some measures than the (s, a)-rect-MMDP. When the parameters are closely concentrated ($\alpha = 100$), we observe that the WVP, max-min-MMDP the (s, a)-rect-MMDP perform similarly in terms



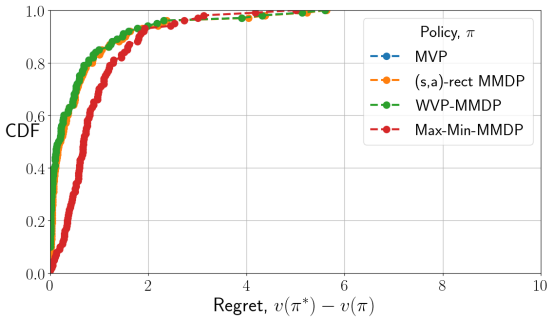
(a) Value function, $\alpha = 10$



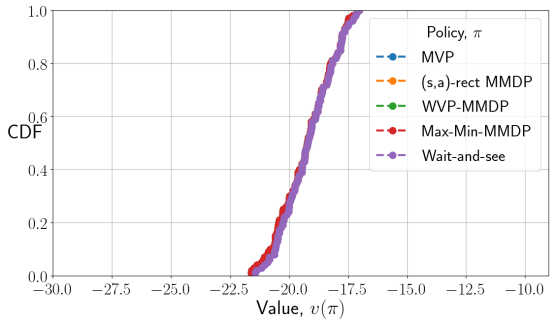
(b) Regret, $\alpha = 10$



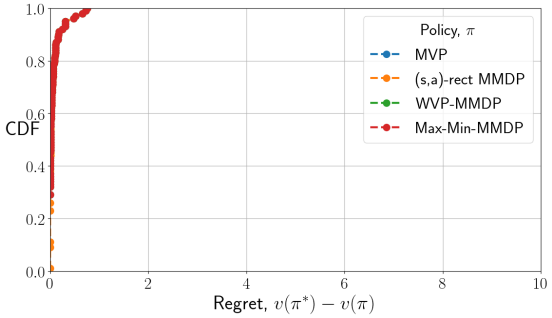
(c) Value function, $\alpha = 20$



(d) Regret, $\alpha = 20$

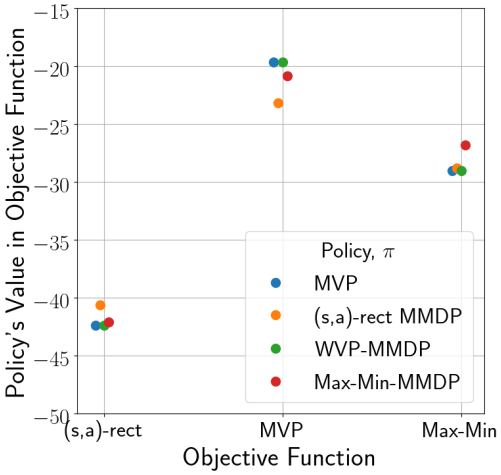


(e) Value function, $\alpha = 100$

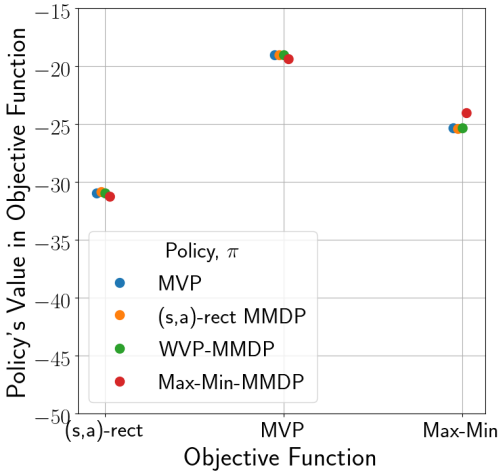


(f) Regret, $\alpha = 100$

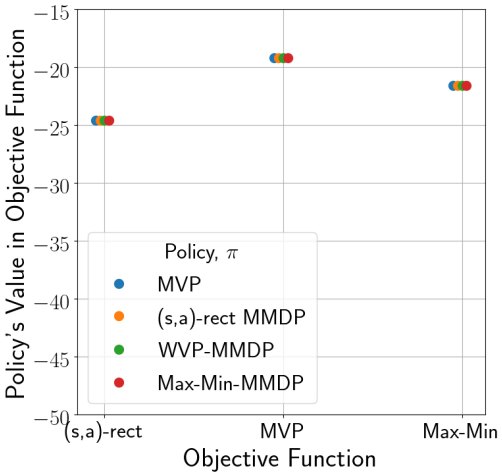
Figure 4.6: CDFs for the value functions and regret corresponding to each of the MMDP policies in the machine maintenance MMDP. Models were generated from a Dirichlet distribution with the corresponding concentration parameters α.



(a) $\alpha = 10$



(b) $\alpha = 20$



(c) $\alpha = 100$

Figure 4.7: A comparison of the MMDP policies as evaluated in the (s, a)-rect-MMDP, MVP-MMDP, and the values in their corresponding worst-case models. Models were generated from a Dirichlet distribution with a concentration parameter with the corresponding concentration parameters, α.

of regret (see Figure 4.6(f)).

Figure 4.7 compares the value functions of these policies in terms of their performance in the (s, a)-rect-MMDP, the MVP-MMDP, and their corresponding worst-case model. We see that the MVP, WVP, and max-min-MMDP do not look like attractive candidate policies when evaluated in the (s, a)-rect-MMDP. The MVP- and WVP-MMDP policies perform about 4.4% worse than the (s, a)-rect-MMDP on this metric and the max-min-MMDP performs 3.6% worse than the (s, a)-rect-MMDP. However, in terms of the actual worst-case performance, the MVP performs within 1% of the (s, a)-rect-MMDP. Therefore, measuring robustness as the performance in the (s, a)-rect-MMDP may not be the best to evaluate the true robustness of a policy if the underlying ambiguity does not satisfy the (s, a)-rectangularity assumption. As the concentration parameter increase from $\alpha = 10$ to $\alpha = 100$, we start to observe that the optimal values for the (s, a)-rect-MMDP and the max-min-MMDP tend towards the MVP. However, even for a very large value of the concentration parameter ($\alpha = 100$) we observe a gap of 3.02 between the (s, a)-rect-MMDP value function and the max-min-MMDP value function.

In summary, the amount of variance in the models' parameters has a large effect on the extent to which the DM's preference towards ambiguity influences the policy. When the model parameters are very similar and exhibit lower variance, the various MMDP recommend similar policies. This finding suggests that the MVP may actually be a policy that is quite robust to small variations. However, as the models' parameters exhibit more variance, we observe that the DM's preferences towards ambiguity are more important to consider as it can influence how the DM should act. We observe that in the case where the model parameters are quite different, the (s, a)-rectangular projection of the MMDP does not provide a good approximation of actual worst-case performance and DM should use caution in employing this property if only for the sake of computational gain.

4.4.2. Case study: Cardiovascular disease management

In this section, we present an MMDP in the context of CVD management. In this study, we seek to optimize the timing of statin initiation and intensification patients with type 2 diabetes, who are at particularly high risk of CVD. The case study presented here is similar to that presented in Chapter 2. However the case study presented here only focuses on cholesterol control and only considers statin therapy. We consider the optimal timing of initiation of a low-dose statin in conjunction with the optimal timing of a high-dose statin.

Intensifying to higher-dose statin has been shown to provide clinical benefit [39]. Below, we use an MMDP model to investigate the timing of initiating and intensifying statins under the presence of ambiguity in the model of TC and HDL progression.

MMDP formulation

The MDP formulation of Mason et al. [47] was adapted to create an MMDP to consider low-dose statins and high-dose statins. The state space of the MMDP is a finite set of health states defined by TC and HDL, and whether or not the patient is currently on no medication, a low-dose statin, or a high-dose statin. A discrete set of actions represent the initiation and intensification of statin therapy. The objective is to optimize the timing of statin initiation and intensification to maximize a weighted combination of QALYs and monetary costs. Figure 4.8 provides a simplified example to illustrate the problem for the MDP used to model this decision process. In the diagram, solid lines illustrate the possible transitions among the cholesterol states and the state representing an adverse event (stroke or CHD event), or death from other causes. In each medication state, including the no medication state ($\emptyset$), patients probabilistically move between health risk states, represented by L (low), M (medium), H (high), and V (very high).

Treatment *actions* are taken at a discrete set of decision epochs indexed by $t \in \mathcal{T} = \{0, 1, \ldots, T\}$ that correspond to ages 40 through 74 at one-year intervals that represent annual preventive care visits with a primary care doctor. States can be separated into *living states* and *absorbing states*. Each living state is defined by the factors that influence a patient's cardiovascular risk: the patient's TC and HDL levels and medication state. We denote the set of the TC states by $\mathcal{L}_{\text{TC}} = \{L, M, H, V\}$, with similar definitions for HDL, $\mathcal{L}_{\text{HDL}} = \{L, M, H, V\}$. The thresholds for these ranges are based on established clinically-relevant cut points for treatment [23]. The dashed lines indicate that the DM has decided to initiate a low-dose statin for a patient who is currently not taking a statin, or a high-dose statin indicating that the patient's treatment should be intensified by starting a high-dose statin.

The set of medication states is $\mathcal{M} = \{(LD, HD) : (LD, HD) \in \{0, 1\}^2, LD + HD \leq 1\}$ which corresponds to the ability to be on a low-dose statin (LD), a high-dose statin (HD), or neither, but not both. For a medication state τ and $i \in \{LD, HD\}$, if $\tau_i = 0$, the patient is not on a statin of type i, and if $\tau_i) = 1$, the patient is on a statin of dose type i. The treatment effects for statin dose i are denoted by $\omega^{\text{TC}}(i)$, for the proportional reduction

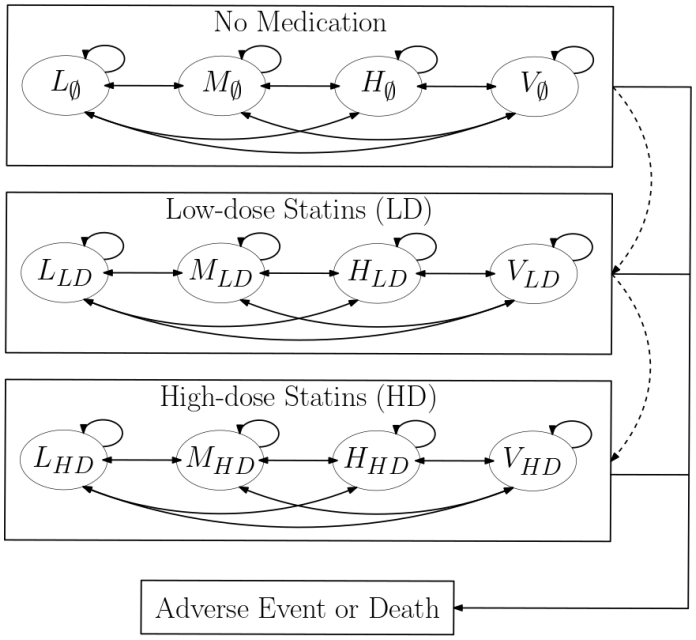


Figure 4.8: A stylized diagram of possible transitions in the CVD MMDP. Solid lines represent possible transitions while dashed lines represent transitions that only occur if an initiation action is taken.

in TC and $\omega^{\text{HDL}}(i)$, for the proportional change in HDL. The low-dose statin effects are as reported in Mason et al. [47], and for the purposes of this case study, we assume that high-dose statins change cholesterol levels by an additional 50% over the change due to low-dose statins. The living states in the model are indexed by $(\ell, \tau) \in \mathcal{L} \times \mathcal{M}$. The absorbing states are indexed by $d \in \mathcal{D} = \{\mathcal{D}_{\text{S}}, \mathcal{D}_{\text{CHD}}, \mathcal{D}_{\text{O}}\}$ represent having a stroke, $\mathcal{D}_{\text{S}}$, having a CHD event, $\mathcal{D}_{\text{CHD}}$, or dying, $\mathcal{D}_{\text{O}}$. The action space depends on the history of medications that have been initiated in prior epochs such that patients can must start a low-dose statin at least a year before intensifying to a high-dose statin. Depending on the current medication state, the action space could consist of initiating a low-dose statin (LD), intensifying to a high-dose statin (HD), or wait to initiate or intensify (W):

$$A_{(\ell,\tau)} = \begin{cases} \{LD, W\} & \text{if } \tau = (0,0), \\ \{HD, W\} & \text{if } \tau = (1,0), \\ \{W\} & \text{if } \tau = (0,1), \end{cases}$$

If a patient is in living state (ℓ, τ) and takes action a, the new medication state is denoted by τ', where τ'_i is set to 1 for statin dose i that are newly initiated by action $\mathbf{a}$; $\tau'_i = \tau_i$ for all medications i which are not newly initiated. Once medication i is initiated, the associated risk factor is modified by the medication effects denoted by $\omega^{\text{TC}}(i)$ or $\omega^{\text{HDL}}(i)$, resulting in a reduction in the probability of a stroke or CHD event.

Two types of transition probabilities are incorporated into the model: probabilities of transition among health states and the probability of events (fatal and nonfatal). At epoch t, $\bar{p}_t^{\tau}(d|\ell)$ denotes the probability of transition from state $(\ell, \tau) \in \mathcal{L} \times \mathcal{M}$ to an absorbing state $d \in \mathcal{D}$. Given that the patient is in health state $\ell \in \mathcal{L}$, the probability of being in health state ℓ' in the next epoch is denoted by $q_t(\ell'|\ell) = q^{HDL}(\ell'|\ell) \cdot q^{TC}(\ell'|\ell)$, where $q^{HDL}(\ell'|\ell)$ describe the probabilities of transitioning from the corresponding HDL states of ℓ to ℓ'. The value $q^{TC}(\ell'|\ell)$ is analogous. The mean values of the health state transition probabilities, $\hat{q}^{HDL}(\ell'|\ell)$ and $\hat{q}^{TC}(\ell'|\ell)$ were computed from empirical data for the natural progression of cholesterol adjusted for the absence of medication [20]. We define $\bar{p}_t(j|\ell, \tau, a)$ to be the probability of a patient transitioning to an absorbing state $d \in \mathcal{D}$ given their current state (ℓ, τ) and the medication decisions a. The transition probabilities can be written as:

$$p_t((\ell', \tau')|(\ell, \tau), a) = \begin{cases} \big(1 - \sum_{d \in \mathcal{D}} \bar{p}_t^{\tau}(d|\ell)\big) q^{HDL}(\ell'|\ell) q^{TC}(\ell'|\ell) f(\tau, \tau', a) & \text{if } \ell, \ell' \in \mathcal{L} \\ & \tau, \tau' \in \mathcal{M}, \\ \bar{p}_t(j|\ell, \tau, a) & \text{if } i \in \mathcal{L},\ j \in \mathcal{D}, \\ 1 & \text{if } i = j \in \mathcal{D}, \\ 0 & \text{otherwise.} \end{cases}$$

In the definition above, $f(\tau, \tau', a)$ is an indicator function representing which takes on a value of 1 if the transition from a state with medications τ to a state with medications τ' is possible given the action describe by a, and a value of 0 otherwise.

In contrast to the case study in Chapter 2, we assume that the risk function describing transitions to the absorbing states is known and that the ACC/AHA study represents the risks $\bar{p}_t^{\tau}(d|i)$ for $i \in \mathcal{L}, d \in \mathcal{D}$. We estimate all other cause mortality using the Centers for Disease Control and Prevention life tables [3].

The reward $r_t(\ell, \tau)$ for a patient in health state ℓ at epoch t is:

$$r_t(\ell, \tau) = w\mathcal{Q}(\ell, \tau) - c(\tau),$$

where $\mathcal{Q}(\ell, \tau) = 1 - d^{\text{MED}}(\tau)$ is the reward for one QALY, $w \in \mathbb{R}$ is the *willingness-to-pay* ratio, and $c(\tau)$ is the cost of taking medications τ. QALYs are elicited through patient surveys, and are commonly used for health policy studies [29]. The *disutility* factor, $d^{\text{MED}}(\tau)$, represents the estimated decrease in quality of life due to the side effects associated with the medications in τ. To reflect a strong disutility associated with taking statin therapy, we use a utility decrement of 0.01 for low-dose statins and 0.015 for high-dose statins. These utility decrement values are higher than those used in Chapter 2 and were selected for the purpose of illustrating the impact of ambiguity. We use the costs reported in Mason et al. [47]. The willingness-to-pay value $w \in \mathbb{R}$ is used to represent a societal perspective wherein society is willing to pay w for one QALY [57].

Model generation In this MMDP, we consider how variation in the estimates q^{TC} and q^{HDL} affects the optimal timing of statin initiation and intensification. To understand the impact of ambiguity on the performance of these, we once again use a Dirichlet distribu-

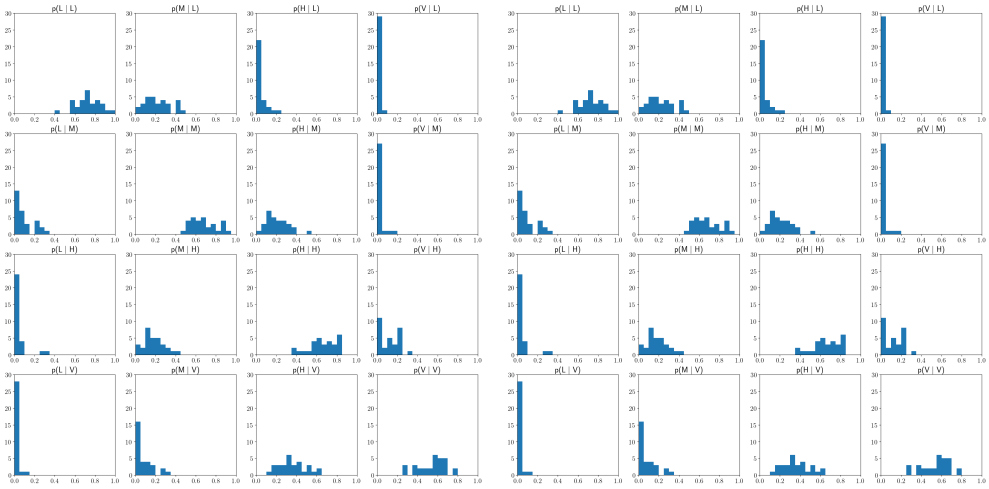
tion. For each row $j \in \{\text{L, M, H, V}\}$, we draw 30 samples of this row from the Dirichlet distribution with parameters $(\alpha q(L|j), \alpha q(M|j), \alpha q(H|j), \alpha q(V|j))$. As before, the concentration parameter α is varied as a way to control the amount of ambiguity in the MMDP for our experiments. We consider values $\alpha = 10$, and 20. Histograms of each element in the transition matrix can be found in Figure 4.9 illustrating the level of ambiguity in the transition probability parameters. Each model is given equal weight in the MMDP.

Experiments We solved the MMDPs described above using the modified B&B algorithms and the rectangular projection using Algorithm 6 for a willingness-to-pay of \$20,000 per QALY. \$20,000 per QALYs was selected for the purposes of illustration and does not necessarily reflect the norm in public health studies. We consider this willingness-to-pay value because this is a particular instance of the MMDP under which ambiguity has a significant effect. For higher values of the willingness-to-pay, statins are more commonly initiated and intensified, and the irreversible nature of these decisions causes ambiguity to affect the value functions only rather than the recommended course of treatment. We analyzed statin therapy for women because our initial experiments showed that the timing of statin initiation in women is more sensitive to ambiguity than men, and thus it presents a more compelling case study for our purposes. The increased sensitivity of womens' initiation of statins might be explained by their tendency to have lower risk than men, perhaps making the decision more dependent on how their disease might progress. For men, it is more clear that there would be a benefit to starting statins, even if the exact transition dynamics are not known precisely.

For each MMDP formulation, we solved both instances within an optimality gap of 2%. If the algorithm did not terminate after 300, we report their gaps at 300 seconds. All instances were solved on a Macintosh MacBook Pro with a 2.7 GHz CPUs and 16 GB of memory using the B&B algorithm implemented in C++ using Xcode Version 10.0. In this case, all MMDP formulations for both $\alpha = 10$ and $\alpha = 20$ were solved in 0.3 seconds except for the min-max-regret-MMDP formulation which had a gap of 100% in both cases after 300 seconds.

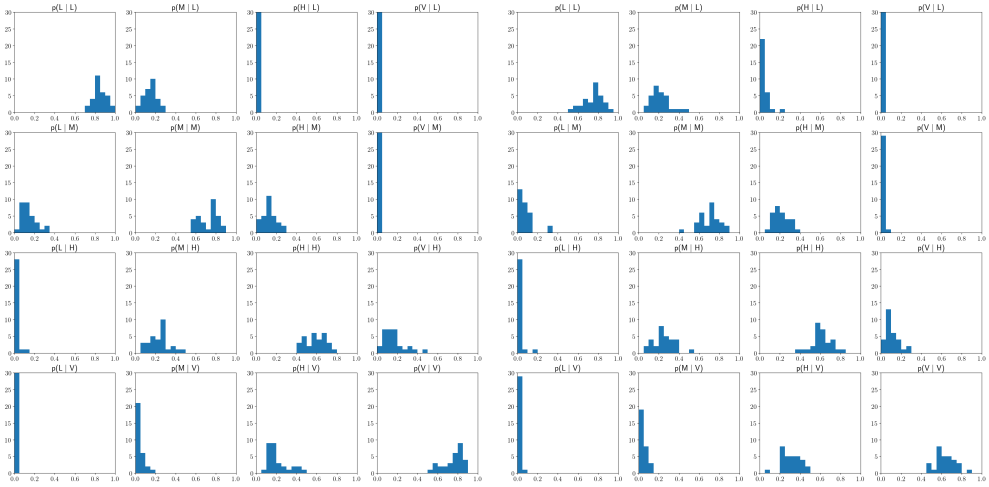
Results

Policy comparisons We now present our findings for the case where the model parameters are more closely concentrated around their mean values. Figure 4.10(a) shows the



(a) HDL, $\alpha = 10$

(b) TC, $\alpha = 10$



(c) HDL, $\alpha = 20$

(d) TC, $\alpha = 20$

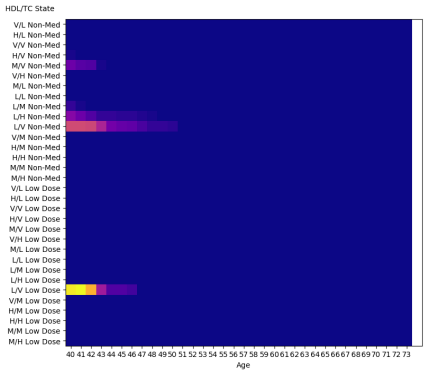
Figure 4.9: Histograms of samples from the Dirichlet distributions used in the CVD MMDP. Each figure corresponds to a histogram of the 30 samples from the corresponding Dirichlet distributions used to construct the models in the CVD MMDP for both the HDL transition matrix and the TC transition matrix. Higher values of the concentration parameter give samples that are more closely concentrated around their mean values.

optimal policies for the "wait-and-see" problem which the optimal policies for each model superimposed on top of each other to create a heat-map. The lightest portions of the policy represent state-time pairs for which all models agree that the patient should initiate or intensify statin therapy in these states. The darkest portions of the policy show the state-time pairs for which all models agree that the patient should defer the initiation or intensification of statin therapy for at least another year. Thus, the figure illustrates that ambiguity in the transition dynamics gives rise to ambiguity as far as the best time to initiate statins for most female patients with low levels of HDL and very high levels of TC. Further, it is unclear whether patients who develop extremely bad cholesterol levels between ages 42-50 should intensify their statin dose.

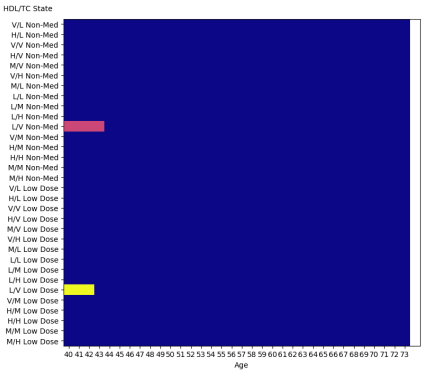
Figures 4.10(b)-4.10(f) illustrate the recommended policies based on the policies for the MVP, WVP, PercOpt, max-min, and (s, a)-rect-MMDPs. The pink portions represent the cholesterol levels and ages for which patients that are not currently taking statins should initiate low-dose statin therapy. The yellow portions represent the cholesterol levels and ages for which patients that are already taking a low-dose statin should intensify to high-dose statin therapy. Notice that, besides the (s, a)-rect-MMDP and the MVP-MMDP, all of the ambiguity-aware MMDPs provide the same recommendations. The MVP-MMDP policy differs only in its suggestion for women with extremely bad cholesterol and aged 45 to not initate statin therapy. The other MMDP policies suggest that the patient should only start statins if their cholesterol is very bad at a young age (low HDL and very high TC, age 45 or younger). The (s, a)-rect-MMDP is more aggressive in the initiation and intensification of statin therapy. The policy recommends statin intiation and intensification for more moderate levels of cholesterol and for older patients. Others who have considered (s, a)-rectangular ambiguity sets for medical decision making have also described this phenonmenon. Zhang, Steimle, and Denton [80] observe that a robust MDP approach tends to initiate second-line medications for glycemic control sooner than its nominal counterpart. Kaufman, Schaefer, and Roberts [37] also show that a robust MDP model of liver transplantation suggests earlier therapy than the nominal MDP formulation. Sinha, Kotas, and Ghate [66] show that dosing in a robust MDP formulation with an (s, a)-rectangular interval model ambiguity set recommends higher doses than its nominal counterpart. Our results show these robust formulations based on (s, a)-rectangular projections of ambiguity tend to give more aggressive treatment than those that do not do this projection. Figure 4.12(a) illustrates the policies when the models of the MMDP are sampled from a Dirichlet distribution with concentration parameter $\alpha = 10$. Even with more variation in the tran-

sition dynamics, the MMDP policies (except for (s, a)-rect-MMDP) all recommend the same treatment strategy while (s, a)-rect-MMDP initiates and intensifies statin therapy more aggressively.

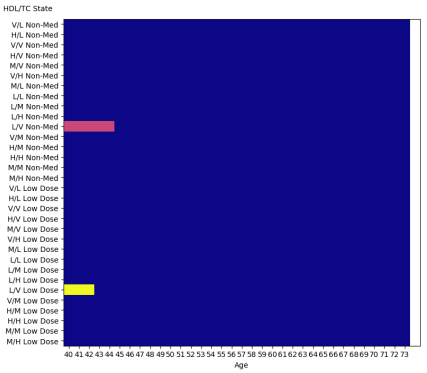
Figure 4.11(a) shows the optimal policies when the model parameters are more closely concentrated around their mean values. The "wait-and-see" problem shows that the ambiguity in the transition dynamics gives rise to ambiguity as far as the best time to initiate statins for most female patients with low levels of HDL and very high levels of TC. Further, it is unclear whether patients who develop extremely bad cholesterol levels between ages 42-47 should intensify their statin dose. Figures 4.11(b)-4.11(f) illustrate the recommended policies based on the policies for the MVP, WVP, PercOpt, max-min, and (s, a)-rect MMDPs. Notice that in this case, besides the (s, a)-rect-MMDP, all of the ambiguity-aware MMDPs provide the same recommendations. These policies suggest that the patient should only start statins if their cholesterol is very bad at a young age (low HDL and very high TC, age 45 or younger). Once again, the (s, a)-rect-MMDP is more aggressive in the initiation and intensification of statin therapy, although not to quite the same extent.



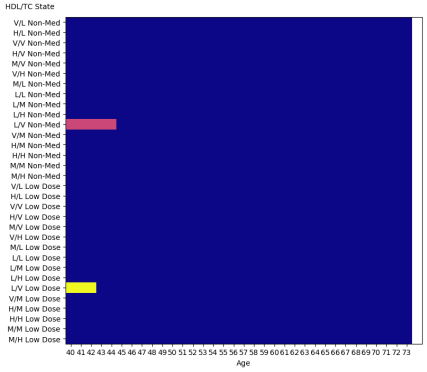
(a) Heatmap of wait-and-see



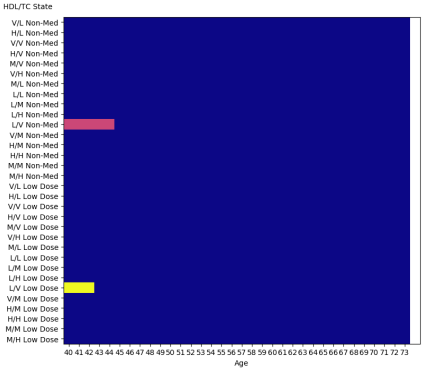
(b) The mean value problem policy



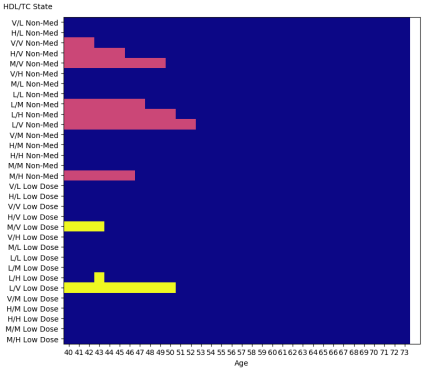
(c) The weighted value policy



(d) Perc-Opt-MMDP (20%) policy

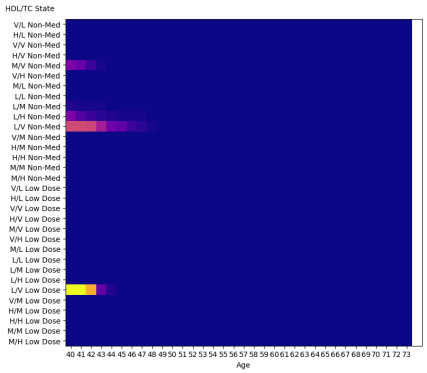


(e) The max-min MMDP policy

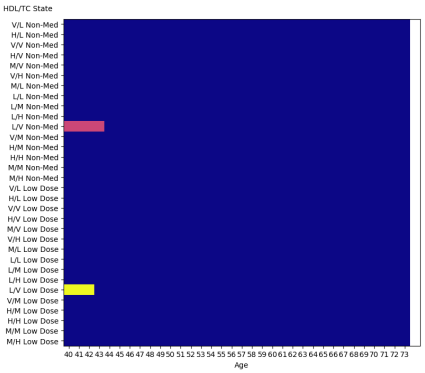


(f) The (s, a)-rectangular MMDP policy

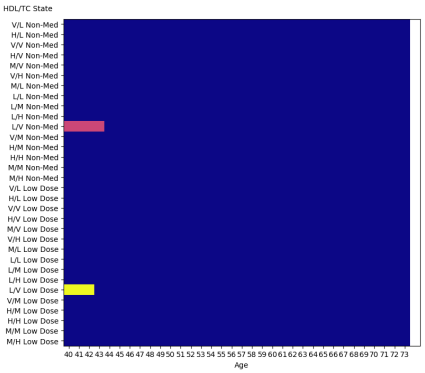
Figure 4.10: An illustration of the ambiguity-aware MMDP policies for the various formulations of the cardiovascular disease management MMDP when the Dirichlet distribution has a concentration parameter of $\alpha = 20$. Blue represents that the optimal action for this health state and age is to wait to defer initiation of intensification of a statin. Pink (yellow) represents that the optimal action is to initiate (intensify) a statin. For the wait-and-see policy, the bright pink (yellow) represents states and ages for which all models agree that initiating (intensifying) a statin is the optimal decision. The darker shades represent states for which the models' policies disagree in terms of the optimal action.



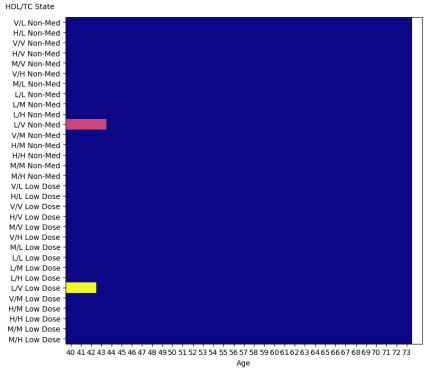
(a) Heatmap of wait-and-see



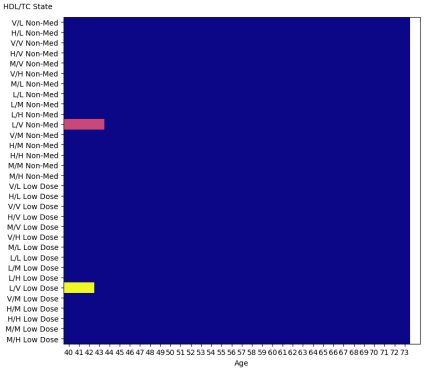
(b) The mean value problem policy



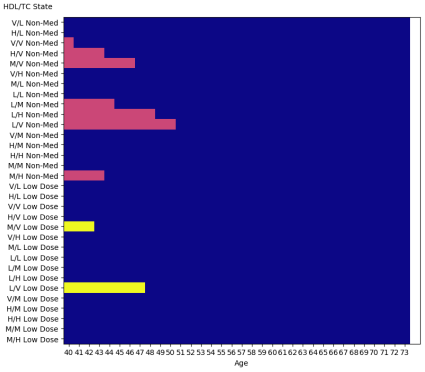
(c) The weighted value policy



(d) Perc-Opt-MMDP (20%) policy



(e) The max-min MMDP policy



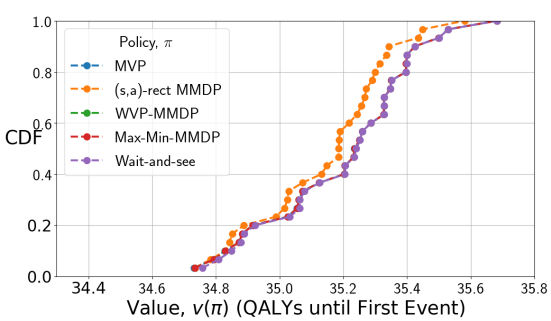
(f) The (s, a)-rectangular MMDP policy

Figure 4.11: An illustration of the ambiguity-aware MMDP policies for the various formulations of the cardiovascular disease management MMDP when the Dirichlet distribution has a concentration parameter of $\alpha = 20$. Blue represents that the optimal action for this health state and age is to wait to defer initiation of intensification of a statin. Pink (yellow) represents that the optimal action is to initiate (intensify) a statin. For the wait-and-see policy, the bright pink (yellow) represents states and ages for which all models agree that initiating (intensifying) a statin is the optimal decision. The darker shades represent states for which the models' policies disagree in terms of the optimal action.

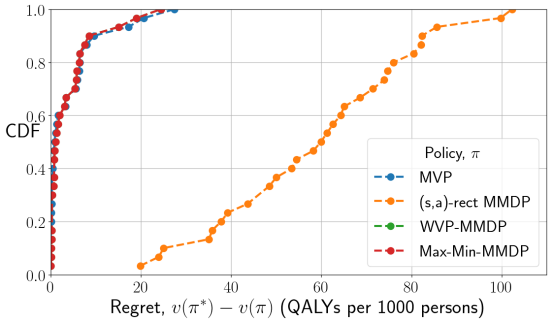
Value function comparisons Figure 4.12(a) and 4.12(a) illustrates the CDF of the value function corresponding to the MMDP policies for $\alpha = 10$ and $\alpha = 20$ respectively. For $\alpha = 10$, we see that the (s, a)-rect-MMDP achieves a lower mean than the other MMDP policies. The (s, a)-rect-MMDP achieves approximately the same as the MVP policy in their respect worst-cases (34.73 expected QALYs before first event). However, the CDF of the MVP nearly dominates the CDF of the (s, a)-rect-MMDP. We see that the MVP also performs well on other metrics such as the 0.2-percentile, best worst-case performance, and best weighted value. The wait-and-see line shows an optimistic CDF that would come from solving each model of the MMDP independently. For $\alpha = 20$, we see that the variance in the value function estimates is lower relative to the case with a higher level of dispersion in the model parameters. In this figure, we see that the (s, a)-rect-MMDP still performs worse than the other MMDP policies. The (s, a)-rect-MMDP achieves approximately the same as the MVP policy in their respect worst-cases (34.81 expected QALYs before first event). Once again, the CDF of the MVP nearly dominates the CDF of the (s, a)-rect-MMDP. We see that the MVP also performs well on other metrics such as the 0.2-percentile, best worst-case performance, and best weighted value.

Figure 4.12(d) illustrates these differences as measured by regret in terms of QALYs per 1000 persons. We see that projecting the MMDP onto a (s, a)-rectangular ambiguity set and solving (s, a)-rect-MMDP can lead to a policy that underperforms with respect to each individual model. For $\alpha = 10$, the worst-case regret for the (s, a)-rect-MMDP is 102.3 QALYs per 1000 persons while the MVP achieves a worst-case regret of 27.5 QALYs per 1000 persons. Therefore, although the (s, a)-rect-MMDP aims to be robust to deviations in the model parameters, it actually underperforms in terms of worst-case regret and most other metrics. Comparing model-by-model, we observe that the MVP-MMDP does between 1.4 to 87.5 QALYs per person better than the (s, a)-rect-MMDP's policy. To put these values in perspective, the use of aspirin for secondary prevention of myocardial infarction in 45-year-old men which has been estimated to provide a QALY gain of 40 per 1000 patients [55] and aspirin is considered an important intervention. For $\alpha = 20$, the worst-case regret for the (s, a)-rect-MMDP is 90.3 QALYs per 1000 persons while the MVP achieves a worst-case regret of 20.5 QALYs per 1000 persons. We observe again that although the (s, a)-rect-MMDP aims to be robust to deviations in the model parameters, it actually underperforms in terms of worst-case regret and most other metrics.

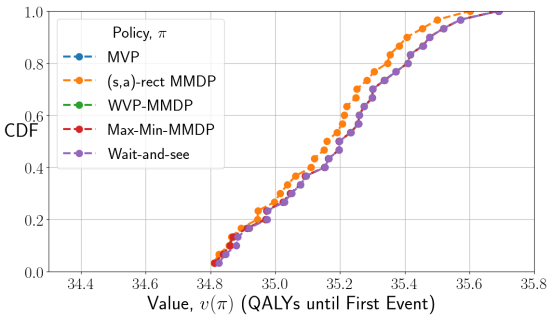
Figure 4.13(a) and 4.13(b) compares the value functions of these policies in terms of their performance in the (s, a)-rect-MMDP, the MVP-MMDP, and their corresponding worst-



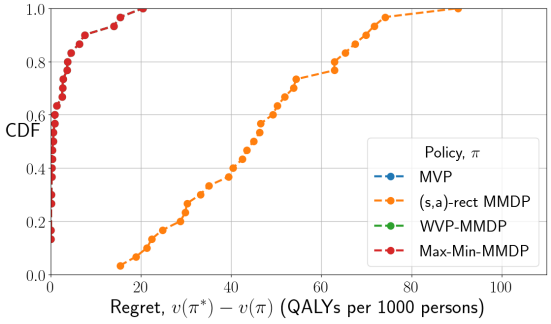
(a) Value functions, More dispersion ($\alpha = 10$)



(b) Regret, More dispersion ($\alpha = 10$)

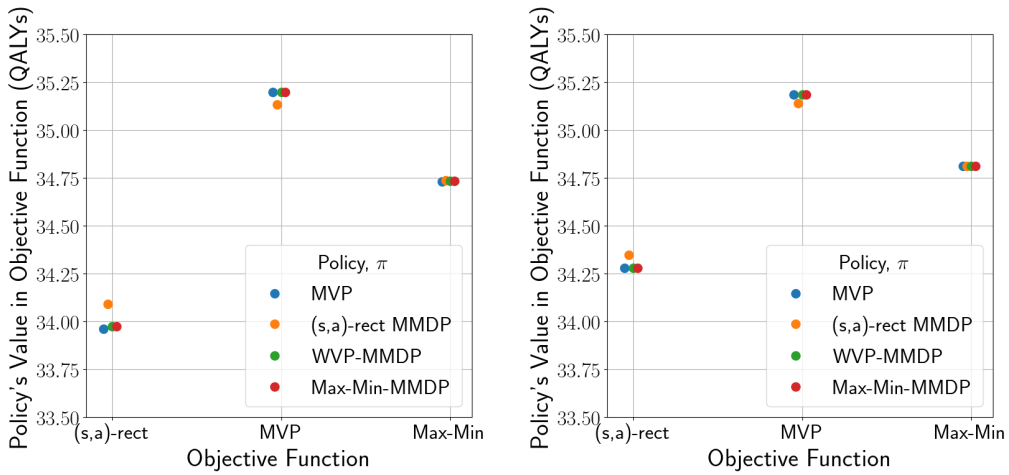


(c) Value functions, Less dispersion ($\alpha = 20$)



(d) Regret, Less dispersion ($\alpha = 20$)

Figure 4.12: CDFs for the value functions and regret corresponding to each of the MMDP policies in the MMDP for CVD management. Models were generated from a Dirichlet distribution with the corresponding concentration parameters α.



(a) $\alpha = 10$ (b) $\alpha = 20$

Figure 4.13: A comparison of the MMDP policies as evaluated in the MVP, the (s, a)-rect-MMDP and their worst-case values. Models were generated from a Dirichlet distribution with the corresponding concentration parameter, α.

case model. As expected, the (s, a)-rect and MVP-MMDP policies achieve the maximum value for their respective problems. For the higher level of dispersion (α=10), the optimal value for the (s, a)-rect-MMDP is 34.1 while the (s, a)-MMDP's worst-case model value is 34.7. Also notice that the MVP performs only slightly worse on the max-min-MMDP as the (s, a)-rect-MMDP policy. For $\alpha = 20$, the optimal value for the (s, a)-rect-MMDP is 34.4 while the (s, a)-MMDP's worst-case model value is 34.8. Once again, we observe that although the MVP policy and the (s, a)-rect-MMDP performs similarly in terms of their worst-case models. This finding suggests that comparing policies on the basis of performance in the (s, a)-rect-MMDP may not be a fair way to compare their robustness if the ambiguity in the MDP's parameters do not satisfy the rectangularity assumption.

4.4.3. Discussion

We now discuss some of the most interesting findings from the case studies. First, we note that the other ambiguity-aware formulations of MMDPs provide policies that might be more appropriate for DMs who are not risk-neutral to ambiguity. There are some cases where the MMDP formulations are especially effective in managing other measures of the impact of ambiguity, such as the worst-case performance among all models. However, we have found that when the transition probability parameters are more closely concentrated

around their mean values (e.g., in instances when the concentration parameter was larger), the optimal solution to the MVP tends to serve as a good policy as evaluated on several alternative ambiguity-aware measures of performance for MMDPs, such as its evaluation in the WVP-MMDP, max-min-MMDP, and min-max-regret-MMDP. In cases where the transition probabilities for the models are quite different, the DM's preference towards ambiguity becomes more important.

Our case studies also illustrated the following trends in the impacts of ambiguity on the optimal policies and value functions. First, we observe that for a fixed policy π, the value function $v(\pi, P)$ tends to be sensitive to the transition probability matrix, P. We also observe that the policy that optimizes the value function $v(\pi, P)$ is sensitive to changes in P; however, we see that the implications of this can be minor. For instance, suppose that $\pi(P)$ represents the policy that maximizes $v(\pi, P)$. We observe that for many test cases $v(\pi(P^1), P^1) - v(\pi(P^2), P^1)$ tends to be small if P^1 is "close to" P^2. This gives some explanation as far as why the solution to the MVP is a near-optimal solution in many models when the MVP's parameters are close to the individual models' parameters. Future analysis may consider bounds on the loss of the MVP. Some work has considered this in the case of MDPs with ambiguous transition matrices in an adversarial setting [48], but the view of MMDP models as random variables with finite support may allow for tighter bounds.

We found that in many cases the (s, a)-rect-MMDP performs worse than all other policies on several ambiguity-aware measures of performance in the models, which has been the focus in this thesis. We also found that although the (s, a)-rect-MMDP is supposed to protect against deviations, the projection of the MMDP onto a (s, a)-rectangular ambiguity set may not be a good approximation of the actual non-rectangular ambiguity. We see that, in some cases, using the (s, a)-rectangular projection can perform worse in terms of regret than the MVP.

Further, the value function in the (s, a)-rect-MMDP may severely underestimate the value function of the worst-case model in the MMDP for specific policies. We find that comparing on the basis of performance in the (s, a)-rect-MMDP may not be an accurate reflection of the actual worst-case performance. This finding can have important implications, as seen in the CVD management case study. As Zhang, Steimle, and Denton [80] and Kaufman, Schaefer, and Roberts [37] have observed in other studies on medical decision making under ambiguity, the (s, a)-rect-MMDP in the CVD case study produces policies that are extremely aggressive in terms of the initiation of treatment. However, we

observe that these policies perform strictly worse than the MVP when evaluated in the actual MMDP models. We observe that as the model parameters become more closely concentrated, the impact of this assumption is not as extreme. It could be beneficial to DMs to determine some measure of violation of the (s, a)-rectangularity assumption, which could provide DM's with an understanding of how well the (s, a)-rectangular projection approximates the actual ambiguity in the MDP.

4.5. Conclusion

In this chapter, we presented a collection of ambiguity-aware formulations of the non-adaptive MMDP that modify objective functions to provide a suitable policy to the DM who is risk-averse to ambiguity in the model's parameters. We presented the max-min MMDP wherein the DM sought to maximize the performance in the worst-case realization of the model, the min-max-regret MMDP wherein the DM sought to minimize the maximum regret between the value of the policy used in a model and the best possible value achievable under that model's parameters. Finally, we presented the PercOpt-MMDP which optimizes for the ϵ-percentile.

We provided the definitions of these formulations, summarizing existing complexity results and solution methods. The scenario-based approaches tend to be NP-hard, and existing solution methods have been limited to MIP approaches. We showed that these ambiguity-averse MDPs can also be solved using modifications to the B&B algorithm used to solve the non-adaptive WVP which can also be used to solve these alternative formulations of the MMDP. However, the degree of difficulty in solving these formulations varied considerably, with the min-max-regret and PercOpt MMDPs being most difficult to solve. Thus, providing a better warm-start to the min-max-regret-MMDP B&B algorithm and further study of branching and node selection strategies to more quickly raise the lower bound is necessary could be an important future direction for research.

We illustrated these formulations on two case studies: one in the context of machine maintenance and the other in the context of cholesterol management for the prevention of CVD for type 2 diabetes patients. Our findings illustrated that in some cases, the alternative preferences might influence the DM's best course of action, but in many cases, the MVP provides a solution that performs well for the various objective functions. We also compared our ambiguity-aware MMDP formulations to a robust MDP approach wherein

the MMDP's parameters are projected onto an ambiguity set that would satisfy the commonly employed (s, a)-rectangularity property. We showed that enforcing this rectangularity assumption when it is not appropriate could lead to a policy that under-performs if the underlying ambiguity does not satisfy this property. We showed that in some cases, a DM would be better off simply solving the MVP rather than employing rectangularity for the sake of tractability and that comparing on the basis of the performance in the (s, a)-rectangular projection may not be a measure of actual worst-case performance if the true ambiguity in the model does not satisfy the (s, a)-rectangularity assumption.

Our study has the following limitations. First, in our case studies, we use the Dirichlet distribution to generate the models of the transition probability matrices. We did this as a way to control the concentration of the model parameters around their mean and in doing so, we sought to control the amount of ambiguity in the underlying MDP's parameters. However, this was done for the sake of illustration and it may not be the best way for a DM to generate models of the ambiguity in practice. Instead, the DM may want to take another approach. For instance, if the DM is concerned about ambiguity in parameters arising from statistical uncertainty one could use bootstrapping to estimate simultaneous confidence intervals [26] and use a sampling method to draw samples from within the confidence region [81]. Our results suggest that barring very high variation in the resulting parameter estimates, the solution the MVP problem is likely to perform quite well under multiple criteria.

Another limitation is that we compare our policies in terms of the same models used in the optimization process. Future work may consider how best to partition models into a training and validation set of models along the lines of the procedure done in Mannor et al. [46] to better understand the performance of the MMDP policies. Second, we consider only two case studies related to MDPs both of which concern deteriorating systems. We found that the MVP tends to be a near optimal solution under many preferences towards ambiguity, but it may be that the structure of these problems is causing this phenomenon. Third, we limit our analysis to the class of Markov deterministic policies. In Chapter 2, we showed that there will always exist a deterministic policy that is optimal for the non-adaptive MMDP, but it may be that randomized policies may be optimal for other formulations of the non-adaptive MMDP. However, in many practical problems, Markov deterministic policies are optimal because they are transparent and practical from a managerial perspective.

We compared policies resulting from the MMDP to policies resulting from the projection

of the MMDP onto a (s, a)-rectangular ambiguity set. In some cases the (s, a)-rect-MMDP performed similarly to the MVP and other MMDP formulations. However, we showed that in some cases, projecting the MMDP onto a (s, a)-rectangular ambiguity set for the sake of tractability can lead to undesirable outcomes if the (s, a)-rectangularity assumption is not a reasonable assumption. Therefore, we would recommend that the (s, a)-rectangularity assumption is met before employing this assumption simply for computational gain. Future work could investigate how a projection of the MMDP onto another type of ambiguity set could fare in terms of its performance, especially given advances in solving robust MDPs with other ambiguity sets that aim to limit the overly protective nature of the resulting policy (e.g., s-rectangular in Wiesemann, Kuhn, and Rustem [74], k-rectangular in Mannor, Mebel, and Xu [45], and r-rectangular in Goyal and Grand-Clement [30])).

In summary, we compared the existing formulations of MMDPs that are sensitive to ambiguity. We showed that in some cases these preferences could alter the DM's preferred course of action and, if so, the DM is better off using these approaches in contrast to the more computationally attractive robust MDP approach for problems where the (s, a)-rectangularity assumption is not appropriate. However, we find that in many cases, the MVP performs quite well on many instances for multiple objective functions that consider ambiguity.

Chapter 5

Summary and Conclusion

The overall objective of this thesis was to better understand the impact of ambiguity on stochastic dynamic optimization methods and to design new methods to make policies more resilient to ambiguity. Markov decision processes are a very useful mathematical tool for informing sequential decision making under uncertainty, but this model's usefulness can be hindered when there is ambiguity in the underlying stochastic model. To address this problem, we developed the MMDP which allowed us to incorporate ambiguity by using several possible models of the MDP and allowing the DM to consider the performance of a policy with respect to these multiple models. We considered multiple risk preferences towards the characteristics of optimal policies, characterized its complexity, designed new solution methods, and illustrated the utility of these methods on case studies related to CVD management and the repair of machines. Although our methods were applied to these specific decision-making problems in health care and maintenance, our methods extend to other application areas where stochastic dynamic optimization is used, such as finance and inventory management, and transportation systems. Following is a summary of the most important findings of Chapters 2, 3, and 4.

In Chapter 2, we presented a new method for handling ambiguity in MDPs. We introduced the MMDP, which allows for multiple models of the reward and transition probability parameters. We presented the WVP for an MMDP whose solution provides a policy that maximizes the weighted value across multiple models. We identified two important variants of the WVP: the non-adaptive problem and the adaptive problem. We identified important connections between these problems and those in the literature. For example, we showed the adaptive problem is a special case of a POMDP and described solution methods that exploit the structure of the belief space for computational gain. Moreover, we showed that

the non-adaptive problem could be viewed as a two-stage stochastic integer program in which the first-stage decisions correspond to the policy and the second-stage decisions correspond to the value-to-go in each model under the specified policy. This characterization provided insight into a formulation of the non-adaptive problem as an MIP corresponding to the deterministic equivalent problem of the aforementioned two-stage stochastic program. These connections allowed us to quantify the impact of ambiguity on the solution of the MMDP.

We used a case study in the context of CVD management to illustrate the MMDP method. We solved an MMDP consisting of two models which were parameterized according to two well-established but conflicting studies from the medical literature which give rise to ambiguity in the cardiovascular risk of a patient. We used an approximation algorithm to solve the non-adaptive problem addresses this ambiguity by trading off performance between these two models and can achieve a lower expected regret than either of the policies that would be obtained by simply solving a model parameterized by one of the studies, as is typically done in practice currently. For the most part, the policies generated via the WSU approximation algorithm found a balance between the medication usage in each of the models. However, for men, the WSU approximation algorithm suggested that more aggressive use of thiazides and ACE/ARBs would be allow for a better balance in performance in both models. For women, the WSU approximation algorithm generated a policy that is more aggressive in cholesterol control than the FHS model's optimal policy and more aggressive in blood pressure control than the ACC/AHA model's optimal policy.

The main findings of Chapter 2 were that the MMDP is difficult to solve computationally, but its solution can be important in terms of mitigating the impact of ambiguity. Furthermore, a fast polynomial time approximation can provide near-optimal solutions for most problem instances. We also found that VSS is often low but EVPI can be high in some cases. Using a case study of CVD management, we showed that it can be important to address ambiguity in MDPs arising from the existence of multiple models and that our approximation algorithm may provide a policy that outperforms solutions that ignore ambiguity.

In Chapter 3, we addressed the problem of solving the non-adaptive WVP for large MMDPs. We proposed two decomposition methods that leverage the problem structure to solve the MMDP. The first was a B&C algorithm in the vein of the Integer L-Shaped Method for two-stage stochastic integer programming with binary first-stage variables. The B&C algorithm decomposed the extensive form of the MIP into a master problem

involving the binary variables used to encode the policy and $|\mathcal{M}|$ subproblems that evaluate a proposed policy in each model of the MDP. Unfortunately, the extensive form relied on the notorious "big-M"s to enforce logical constraints; the big-Ms led to weak optimality cuts to be added within the B&C procedure. We also proposed a B&B procedure which does not require big-Ms in the formulation. The B&B procedure begins by viewing the MMDP as $|\mathcal{M}|$ independent MDPs. The algorithm began by allowing each model of the MDP to have its own policy and sequentially added requirements that the decision rules for certain state-time pairs must agree in each model.

We presented the first numerical study of exact algorithms for MMDPs for realistic problem instances. To do so, we generated random MMDP instances of a machine maintenance problem to compare the computation time required to solve these problems using the extensive form, the B&C procedure, and the B&B procedure. Our computational experiments showed that the B&B solution method greatly outperforms the solution of the extensive form directly and with a B&C method. The B&B solution methods outperform both the extensive form and the B&C on all of our test cases. In general, higher solution times for the B&B procedure resulted when there was higher variance among the models and when there are more models in the MMDP. Our solution methods enabled us to investigate the impact of ambiguity on MDPs. For the machine maintenance instances, we considered the impact of the concentration of the transition probability models around their mean as well as the number of models used in the MMDP on the value of the MMDP approach in terms of value relative to the MVP and expected regret relative to each model's optimal policy. We found that the MMDP approach was most beneficial when the DM has models that are quite different. When the models are similar, the MVP served as a good approximation, but the B&B procedure typically also solved the MMDP quickly in these cases.

The main findings of Chapter 3 suggested that the B&B procedure is a much more promising solution method for solving MMDPs than solution methods that rely on the MIP formulation. With the new ability to solve larger MMDPs, we found that in many instances, the MVP is a near-optimal solution to the WVP, especially when the variance among model parameters tends to be low. In general, our findings suggest that MDPs are often naturally resilient to ambiguity.

In Chapter 4, we presented other ambiguity-aware formulations of the non-adaptive MMDP that modify objective functions to reflect that the DM may not be risk-neutral to ambiguity in the MMDP. The alternate ambiguity-aware preferences included the max-min MMDP, the min-max-regret-MMDP, and the PercOpt-MMDP. In the max-min MMDP,

the DM sought to choose the policy that maximizes the value function for the worst-case realization of the model. In the min-max-regret MMDP, DM sought to find a policy that minimizes the maximum regret as measured by the difference between the optimal value in the model and the value achieved by the policy. Finally, we presented the PercOpt-MMDP in which the DM optimized for the ϵ-percentile.

We provided the definitions of these formulations, summarized existing complexity and solution methods. We noted that the scenario-based approaches tend to be NP-hard and existing solution methods have been limited to MIP approaches. We showed that these ambiguity-averse MDPs can also be solved using modifications to the B&B algorithm presented in Chapter 3.

We used the B&B solution method to solve these formulations for MMDPs in two case studies: one in the context of machine maintenance and the other in the context of cholesterol management for the prevention of CVD for type 2 diabetes patients. Although the B&B algorithm was able to solve the WVP-MMDP and max-min-MMDP, the min-max-regret-MMDP and the PercOpt-MMDP had longer run-times. This suggests that future research could investigate better node selection and branching strategies for these objective functions. Our findings illustrated that in some cases, the alternative preferences might influence the DM's best course of action. In these cases, the DM's preference towards ambiguity shapes the overall distribution of the value functions in the MMDP models. However, in many cases, the MVP provides a solution that performs well for the various objective functions. We also compared our ambiguity-aware MMDP formulations to a robust MDP approach wherein the MMDP's parameters are projected onto an ambiguity set that would satisfy the commonly employed (s, a)-rectangularity property. We showed that employing this rectangularity assumption when it is not appropriate could lead to a policy that under-performs if the underlying ambiguity does not satisfy this property. In the CVD case study, using the (s, a)-rectangular projection of the MMDP led to a policy that was extremely aggressive in terms of the initiation and intensification of statin therapy. We showed that in some cases, a DM would be better off simply solving the MVP rather than employing rectangularity for the sake of tractability and that comparing on the basis of the performance in the (s, a)-rectangular projection may not be a measure of actual worst-case performance if the true ambiguity in the model does not satisfy the (s, a)-rectangularity assumption. In summary, we compared the existing formulations of ambiguity-aware MMDP and showed that they could be incorporated into a common algorithmic framework using minor modifications to the bounding procedure in the B&B.

We illustrated that these formulations might influence the DM's best course of action, but in many cases, the MVP performs well on many metrics related to performance under ambiguity.

The most important findings from Chapter 4 suggested that the B&B framework for solving MMDPs can be easily modified to incorporate other preferences towards ambiguity; however, the computational effort necessary to solved the models varies considerably. Using two case studies, we showed that the MVP can perform well in terms of mitigating the impact of ambiguity with respect to several risk preferences and that a DM should use caution before employing the (s, a)-rectangularity property if it is not a well-supported assumption. Further, we found that the mean of the value function distribution was not sensitive to the DM's preference towards ambiguity. However, the distribution of value functions changed considerably which suggested that the DM might want to consider more than the mean value function in selecting a policy.

There are several interesting extensions of the work presented in Chapters 2, 3, and 4. In Chapter 2, we present an approximation algorithm for the non-adaptive WVP. We analyzed the approximation algorithm and developed bounds. However our lower bound was only for two models. Generalizing the bound to include MMDPs with more than two models could be useful in determining the worst-case performance of this algorithm. Further, we present the adaptive WVP and propose solution methods based on the POMDP literature. However, we only present an exact solution method for the adaptive problem and do not explore approximation algorithms. Although we found that the adaptive solution provided little improvement over the non-adaptive solution on our random test instances, these gains might be amplified over longer time horizons. Therefore, it may be beneficial in some settings to develop approximation methods for the adaptive MMDP.

In Chapter 3, we propose a B&C decomposition in the vein of the Integer L-Shaped Method. We proposed optimality cuts that are based on the dual of the subproblems. We did not explore logic-based optimality cuts which could potentially improve its performance. Further, we did not consider some methods for solving MIPs such as dual decomposition and Lagrangian relaxations. The formulations we provided in Appendix A may provide a starting point for that line of research. Further, these solution methods might be easily modified to handle the infinite-horizon version of an MMDP wherein the DM would like to restrict the possible solutions to the set of stationary deterministic policies. Presumably, the B&B could be modified to solve the relaxation at each node using a standard solution method for infinite-horizon MDPs such as value iteration, policy

iteration, or LP.

In Chapter 4, we considered several existing ambiguity-aware formulations of the MMDP. However, there are other interesting ambiguity-aware formulations that could be desirable from a managerial perspective. For instance, Ghavamzadeh, Petrik, and Chow [24] proposes a safe policy improvement as a way to mitigate ambiguity in an MMDP. They recommend selecting a policy that is guaranteed to provide an improvement over a baseline policy, such as a policy that is already being used in practice. As another measure of robustness, Roy [58] propose (b, w)-robustness in ambiguous environments that are represented by scenarios. In this setting, the DM seeks to find a policy that achieves a performance level of at least b in as many scenarios as possible, but without having a value below w in any scenario. These could be easily incorporated into the objective function of the MMDP and the corresponding solution methods. Our case studies illustrated that in many cases, the MVP performs well on several metrics. Future work could develop bounds on how well the MVP could perform in terms of the performance in the various models. Some work has considered this in the case of MDPs with ambiguous transition matrices in an adversarial setting [48], but the view of MMDPs models as random variables with finite support may allow for tighter bounds. Finally, we showed that in some cases, the (s, a)-rectangular projection of an MMDP does not perform well in terms of the actual worst-case model in the MMDP. It would be worthwhile to develop a measure of measure of violation of the (s, a)-rectangularity property so that DM can better understand if they can employ this assumption for computational gains.

There are many opportunities to build upon the underlying foundations that we have presented. First, applications to CVD management in the medical decision making context demonstrate the potential for application of our approach to other diseases. Our approach could be applied to previously developed models for diabetes, breast cancer, prostate cancer, and many other diseases.

Second, it is an open opportunity to extend the MMDP solution methods to consider more than one policy that must be used in all models. Doing so would allow the DM to determine the added value of allowing for 2, ..., $|\mathcal{M}|$ distinct policies among the $|\mathcal{M}|$ different models and determining which models should be assigned to which policy. This analysis could have important implications in medicine which is moving towards more finely stratified treatment guidelines. Through this lens, we might view each model of the MMDP as the description of the system dynamics for different subpopulations of patients. Generally, one-size-fits-all guidelines are attractive because of their simplicity

and their ease of implementation. On the other extreme, highly stratified guidelines are also desirable because they deliver the most benefit to individual patient groups. Using the MMDP framework, we could vary the number of policies so that policy-makers could better understand the added benefit of tailoring guidelines to subpopulations as the number of policies shifts from one-size-fits-all to one guideline per subpopulation. Further, such an analysis could provide policy-makers with a way to better understand which subsets of patients are most harmed by simple guidelines.

Third, one may extend the MMDP model to consider MDPs that are not all defined on the same state space and action space, which could have many important applications in medicine. It is not uncommon for different research groups to construct mathematical models of a disease that differ in terms of their underlying model structure. The methods we presented in this thesis might be adapted to consider mappings between states and actions of different models.

Another open area of exploration is the investigation of sufficient conditions under which the MMDP's optimal policy has special structure. For instance, it is well known that there are sufficient conditions which guarantee the existence of a monotone policy that is optimal for an MDP. However, it is not immediately clear that if each model of the MMDP satisfies these conditions, that the optimal policy for the MMDP would be monotone. If this were the case, our algorithms could be modified for computational gain.

In conclusion, we investigated models and methods for sequential decision-making under uncertainty in the presence of ambiguity. We presented a model that allows for the DM to identify sets of decisions that will perform well under multiple plausible representations of a system for which there are limitations in the understanding of its dynamics. We studied the structure of these models and designed an algorithmic approach for solving them. As policy-makers are increasingly using mathematical models to inform their recommendations, the work presented in this thesis will provide a framework for decision-making made under ambiguity and serve as an important foundation for future work in stochastic dynamic optimization under ambiguity.

Appendix A

Supplementary material for Chapter 2

In this appendix, we consider alternative formulations of the MMDP which may provide opportunities for future work.

A.1. Another MIP formulation

Note that we can also formulate the MMDP based on the dual of the original MDPs and add non-anticipativity constraints.

$$
\begin{aligned}
\max_{x,\,\pi} \quad & \sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t\in\mathcal{T}} r_t^m(s,a)x_t^m(s,a) + \sum_{s\in\mathcal{S}}\sum_{m\in\mathcal{M}} r_{T+1}^m(s)x_{T+1}^m(s) \\
\text{s.t.} \quad & \sum_{a\in\mathcal{A}} x_1^m(s,a) = \lambda^m \mu_1^m(s), && \forall s\in\mathcal{S}, m\in\mathcal{M}, \\
& \sum_{a\in\mathcal{A}} x_t^m(s,a) - \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} p_{t-1}^m(s|s',a')x_{t-1}^m(s',a') = 0, && \forall s\in\mathcal{S}, t\in\mathcal{T}, m\in\mathcal{M}, \\
& x_{T+1}^m(s) - \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} p_T^m(s|s',a')x_T^m(s',a') = 0, && \forall s\in\mathcal{S}, m\in\mathcal{M}, \\
& x_t^m(s,a) \geq 0, && \forall m\in\mathcal{M}, s\in\mathcal{S}, \\
& && a\in\mathcal{A}, t\in\mathcal{T}, \\
& x_{T+1}^m(s) \geq 0, && \forall m\in\mathcal{M}, s\in\mathcal{S}, \\
& x_t^m(s,a) \leq M\pi_t(a|s), && \forall m\in\mathcal{M}, s\in\mathcal{S}, \\
& && a\in\mathcal{A}, t\in\mathcal{T}, \\
& \sum_{a\in\mathcal{A}} \pi_t(a|s) = 1, && \forall s\in\mathcal{S}, t\in\mathcal{T}, \\
& \pi_t(a|s) \in \{0,1\}, && s\in\mathcal{S}, a\in\mathcal{A}, t\in\mathcal{T}
\end{aligned}
\tag{A.1}
$$

The benefit of this formulation is that all big-M values are upper bounded by M, although tighter bounds may exist.

A.2. Nonlinear formulation of the non-adaptive MMDP

A standard MDP can be solved via the following LP:

$$
(D)\ \max.\ \sum_{s\in S}\sum_{a\in A} r(s,a)x(s,a) \tag{A.2}
$$

$$
\sum_{a\in A} x(s,a) - \sum_{s'\in S}\sum_{a'\in A} \lambda p(s|s',a')x(s',a') = \alpha(s) \qquad \forall s\in S \tag{A.3}
$$

$$
x(s,a) \geq 0\ \ \forall s,a \tag{A.4}
$$

provides insight into the optimal policy. We have that

$$\pi(a|s) = \frac{x(s,a)}{\sum_{a' \in A} x(s,a')} \tag{A.5}$$

Similarly, we can formulate the WVP using mathematical programming. We alter the objective function to represent:

$$\begin{aligned}
\max_{x,\,\pi} \quad & \sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t\in\mathcal{T}} r_t^m(s,a)x_t^m(s,a) + \sum_{s\in\mathcal{S}}\sum_{m\in\mathcal{M}} r_{T+1}^m(s)x_{T+1}^m(s) \\
\text{s.t.} \quad & \sum_{a\in\mathcal{A}} x_1^m(s,a) = \lambda^m \mu_1^m(s), && \forall s \in \mathcal{S}, m \in \mathcal{M}, \\
& \sum_{a\in\mathcal{A}} x_t^m(s,a) - \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} p_{t-1}^m(s|s',a')x_{t-1}^m(s',a') = 0, && \forall s \in \mathcal{S}, t \in \mathcal{T}, m \in \mathcal{M}, \\
& x_{T+1}^m(s) - \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} p_T^m(s|s',a')x_T^m(s',a') = 0, && \forall s \in \mathcal{S}, m \in \mathcal{M}, \\
& x_t^m(s,a) \geq 0, && \forall m \in \mathcal{M}, s \in \mathcal{S}, \\
& && a \in \mathcal{A}, t \in \mathcal{T}, \\
& x_{T+1}^m(s) \geq 0, && \forall m \in \mathcal{M}, s \in \mathcal{S}, \\
& \frac{x_t^m(s,a)}{\sum_{a\in\mathcal{A}} x_t^m(s,a)} - \frac{x_t^{m'}(s,a)}{\sum_{a\in\mathcal{A}} x_t^{m'}(s,a)} = 0, && \forall m, m' \in \mathcal{M}, s \in \mathcal{S}, \\
& && t \in \mathcal{T}, \\
& \sum_{a\in\mathcal{A}} x_t^m(s,a) = 1, && \forall m \in \mathcal{M}, s \in \mathcal{S}, t \in \mathcal{T}, \\
& \pi_t(a|s) \in \{0,1\}, && \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T}
\end{aligned} \tag{A.6}$$

Here, the dual variables x^m correspond to the optimal policy for the corresponding model of the MMDP. That is,

$$\pi_t^m(a|s) = \frac{x^m(s,a)}{\sum_{a'\in\mathcal{A}} x^m(s,a')} \tag{A.7}$$

describes the probability of selecting action a for state s in time t for model m, and the

constraint

$$\frac{x_t^m(s,a)}{\sum_{a\in\mathcal{A}} x_t^m(s,a)} - \frac{x_t^{m'}(s,a)}{\sum_{a\in\mathcal{A}} x_t^{m'}(s,a)} = 0, \forall m, m' \in \mathcal{M}, s \in \mathcal{S}, t \in \mathcal{T} \tag{A.8}$$

ensures that the policies will be the same in all models.

Bibliography

[1] A. Ahmed, P. Varakantham, M. Lowalekar, Y. Adulyasak, and P. Jaillet. "Sampling Based Approaches for Minimizing Regret in Uncertain Markov Decision Processes (MDPs)". In: *Journal of Artificial Intelligence Research* 59 (2017), pp. 229–264.

[2] O. Alagoz, L. M. Maillart, a. J. Schaefer, and M. S. Roberts. "Determining the Acceptance of Cadaveric Livers Using an Implicit Model of the Waiting List". In: *Operations Research* 55.1 (2007), pp. 24–36.

[3] E. Arias and V. Statistics. "National Vital Statistics Reports United States Life Tables , 2007". In: *Statistics* 59.9 (2011), pp. 1–132.

[4] T. Ayer, O. Alagoz, and N. K. Stout. "OR Forum: A POMDP Approach to Personalize Mammography Screening Decisions". In: *Operations Research* 60.5 (2012), pp. 1019–1034.

[5] J. F. Benders. "Partitioning procedures for solving mixed-variables programming problems". In: *Numerische Mathematik* 4 (1962), pp. 238–252.

[6] L. Berger, H. Bleichrodt, and L. Eeckhoudt. "Treatment decisions under ambiguity". In: *Journal of Health Economics* 32.3 (2013), pp. 559–569. ISSN: 0167-6296. DOI: https://doi.org/10.1016/j.jhealeco.2013.02.001. URL: http://www.sciencedirect.com/science/article/pii/S0167629613000088.

[7] D. Bertsimas, D. B. Brown, and C. Caramanis. "Theory and applications of robust optimization". In: *SIAM Review* 53.3 (2011), pp. 464–501.

[8] D. Bertsimas, J. Silberholz, and T. Trikalinos. "Optimal healthcare decision making under multiple mathematical models: application in prostate cancer screening". In: *Health Care Management Science* (Sept. 2016), pp. 1–14. ISSN: 1386-9620.

[9] J. Birge and F. Louveaux. *Introduction to Stochastic Programming*. 1997. ISBN: 0387982175.

[10] R. J. Boucherie and N. M. Van Dijk. *Markov decision processes in practice*. Springer, 2017.

[11] "Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning". In: *Discrete Optimization* 19 (Feb. 2016), pp. 79–102.

[12] A. H. Briggs, A. E. Ades, and M. J. Price. "Probabilistic Sensitivity Analysis for Decision Trees with Multiple Branches: Use of the Dirichlet Distribution in a Bayesian Framework". In: *Medical Decision Making* 23.4 (July 2003), pp. 341–350. ISSN: 0272-989X. DOI: 10.1177/0272989X03255922. URL: http://journals.sagepub.com/doi/10.1177/0272989X03255922.

[13] E. Brunskill and L. Li. "Sample complexity of multi-task reinforcement learning". In: *arXiv preprint arXiv:1309.6821* (2013).

[14] P. Buchholz and D. Scheftelowitsch. "Computation of weighted sums of rewards for concurrent MDPs". In: *Mathematical Methods of Operations Research* (Oct. 2018), pp. 1–42. DOI: 10.1007/s00186-018-0653-1. URL: http://link.springer.com/10.1007/s00186-018-0653-1.

[15] CDC. *2011 Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States*. Tech. rep. U.S. Department of Health, Human Services, Centers for Disease Control, and Prevention, 2011.

[16] G. Codato and M. Fischetti. "Combinatorial Benders' Cuts for Mixed-Integer Linear Programming". In: *Operations Research* (2006).

[17] B. A. Craig and P. P. Sendi. "Estimation of the transition matrix of a discrete-time Markov chain". In: *Health Economics* 11.1 (2002), pp. 33–42.

[18] E. Delage and S. Mannor. "Percentile Optimization for Markov Decision Processes with Parameter Uncertainty". In: *Operations Research* 58.1 (Feb. 2009), pp. 203–213.

[19] Y. Deng, S. Ahmed, and S. Shen. "Parallel Scenario Decomposition of Risk-Averse 0-1 Stochastic Programs". In: *INFORMS Journal on Computing* 30.1 (2017), pp. 90–105.

[20] B. T. Denton, M. Kurt, N. D. Shah, S. C. Bryant, and S. A. Smith. "Optimizing the start time of statin therapy for patients with diabetes". In: *Medical Decision Making* 29.3 (2009), pp. 351–367.

[21] D. Ellsberg. "Risk, Ambiguity, and the Savage Axioms". In: *The Quarterly Journal of Economics* 75.4 (1961), pp. 643–669. URL: http://www.jstor.org/stable/1884324.

[22] R. Etzioni, R. Gulati, A. Tsodikov, E. M. Wever, D. F. Penson, E. A. Heijnsdijk, J. Katcher, G. Draisma, E. J. Feuer, H. J. de Koning, and A. B. Mariotto. "The prostate cancer conundrum revisited: Treatment changes and prostate cancer mortality declines". In: *Cancer* 118.23 (2012), pp. 5955–5963.

[23] E. Expert Panel on Detection and T. of High Blood Cholesterol in Adults. "Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III)". In: *JAMA* 285.19 (May 2001), pp. 2486–2497. ISSN: 0098-7484. DOI: 10.1001/jama.285.19.2486. eprint: https://jamanetwork.com/journals/jama/articlepdf/193847/jsc10094.pdf. URL: https://dx.doi.org/10.1001/jama.285.19.2486.

[24] M. Ghavamzadeh, M. Petrik, and Y. Chow. "Safe policy improvement by minimizing robust baseline regret". In: *Advances in Neural Information Processing Systems*. 2016, pp. 2298–2306.

[25] R. Givan, S. Leach, and T. Dean. "Bounded-parameter Markov decision processes". In: *Artificial Intelligence* 122.1-2 (2000), pp. 71–109.

[26] J. Glaz and C. Sison. "Simultaneous confidence intervals for multinomial proportions". In: *Journal of Statistical Planning and Inference* 82 (Jan. 1997), pp. 251–262. DOI: 10.1016/S0378-3758(99)00047-6.

[27] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. O'Donnell, J. G. Robinson, J. S. Schwartz, S. T. Shero, S. C. Smith, P. Sorlie, N. J. Stone, and P. W. F. Wilson. "2013 ACC//AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology//American Heart Association Task Force on Practice Guidelines". In: *Circulation* 129 (2014), S49–S73.

[28] J. Goh, M. Bayati, S. A. Zenios, S. Singh, and D. Moore. "Data Uncertainty in Markov Chains: Application to Cost-Effectiveness Analyses of Medical Innovations". In: *Operations Research* 66.3 (2018), pp. 697–715.

[29] M. R. Gold, D. Stevenson, and D. G. Fryback. "HALYS and QALYS and DALYS, Oh My: similarities and differences in summary measures of population Health". In: *Annual Review of Public Health* 23.1 (2002), pp. 115–134.

[30] V. Goyal and J. Grand-Clement. "Robust Markov Decision Process: Beyond Rectangularity". In: *arXiv preprint arXiv:1811.00215* (2018).

[31] J. D. F. Habbema, C. B. Schechter, K. A. Cronin, L. D. Clarke, and E. J. Feuer. "Modeling cancer natural history, epidemiology, and control: reflections on the CISNET breast group experience." In: *Journal of the National Cancer Institute. Monographs* 2006.36 (Oct. 2006), pp. 122–126.

[32] A. Hallak, D. Di Castro, and S. Mannor. "Contextual Markov Decision Processes". 2015.

[33] I. Hochberg, G. Feraru, M. Kozdoba, S. Mannor, M. Tennenholtz, and E. Yom-Tov. "Encouraging Physical Activity in Patients With Diabetes Through Automatic Personalized Feedback via Reinforcement Learning Improves Glycemic Control". In: *Diabetes Care* 39.4 (2016), e59–e60. ISSN: 0149-5992. DOI: 10.2337/dc15-2340.

[34] J. N. Hooker and G. Ottosson. "Logic-based Benders decomposition". In: *Mathematical Programming, Series B* (2003). ISSN: 00255610. DOI: 10.1007/s10107-003-0375-9.

[35] G. N. Iyengar. "Robust Dynamic Programming". In: *Mathematics of Operations Research* 30.2 (2005), pp. 257–280.

[36] D. L. Kaufman and A. J. Schaefer. "Robust modified policy iteration". In: *INFORMS Journal on Computing* 25.3 (2013), pp. 396–410.

[37] D. L. Kaufman, A. J. Schaefer, and M. S. Roberts. "Living-Donor Liver Transplantation Timing Under Ambiguous Health State Transition Probabilities". In: *Proceedings of the 2011 Manufacturing and Service Operations Management (MSOM) Conference.* 2011.

[38] G. Laporte and F. V. Louveaux. "The integer L-shaped method for stochastic integer programs with complete recourse". In: *Operations Research Letters* (1993). ISSN: 01676377.

[39] J. C. LaRosa, S. M. Grundy, D. D. Waters, C. Shear, P. Barter, J.-C. Fruchart, A. M. Gotto, H. Greten, J. J. Kastelein, J. Shepherd, and N. K. Wenger. "Intensive Lipid Lowering with Atorvastatin in Patients with Stable Coronary Disease". In: *New England Journal of Medicine* 352.14 (2005). PMID: 15755765, pp. 1425–1435. DOI: 10.1056/NEJMoa050461.

[40] Y. Le Tallec. "Robust, risk-sensitive, and data-driven control of Markov Decision Processes". PhD thesis. Massachusetts Institute of Technology, 2007.

[41] X. Li, H. Zhong, and M. L. Brandeau. "Quantile Markov Decision Process". In: (Nov. 2017). arXiv: 1711.05788.

[42] W. S. Lovejoy. "A survey of algorithmic methods for partially observed Markov decision processes". In: *Annals of Operations Research* 28.1 (Dec. 1991), pp. 47–65. ISSN: 1572-9338. DOI: 10.1007/BF02055574. URL: https://doi.org/10.1007/BF02055574.

[43] J. Luedtke. "A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support". In: *Math. Program., Ser. A* 146 (2014), pp. 219–244. DOI: 10.1007/s10107-013-0684-6.

[44] J. S. Mandelblatt, N. K. Stout, C. B. Schechter, J. J. van den Broek, D. L. Miglioretti, M. Krapcho, A. Trentham-Dietz, D. Munoz, S. J. Lee, D. A. Berry, N. T. van Ravesteyn, O. Alagoz, K. Kerlikowske, A. N. Tosteson, A. M. Near, A. Hoeffken, Y. Chang, E. A. Heijnsdijk, G. Chisholm, X. Huang, H. Huang, M. A. Ergun, R. Gangnon, B. L. Sprague, S. Plevritis, E. Feuer, H. J. de Koning, and K. A. Cronin. "Collaborative Modeling of the Benefits and Harms Associated With Different U.S. Breast Cancer Screening Strategies". In: *Annals of Internal Medicine* 164.4 (Feb. 2016), p. 215.

[45] S. Mannor, O. Mebel, and H. Xu. "Robust MDPs with k-Rectangular Uncertainty". In: *Mathematics of Operations Research* 41.4 (2016), pp. 1484–1509.

[46] S. Mannor, D. Simester, S. Peng, and J. N. Tsitsiklis. "Bias and Variance Approximation in Value Function Estimates." In: *Management Science* 53.2 (2007), pp. 308–322.

[47] J. E. Mason, B. T. Denton, N. D. Shah, and S. A. Smith. "Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients". In: *European Journal of Operational Research* 233.3 (2014), pp. 727–738.

[48] A. Mastin and P. Jaillet. "Loss bounds for uncertain transition probabilities in Markov decision processes". In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. Dec. 2012, pp. 6708–6715. DOI: 10.1109/CDC.2012.6426504.

[49] M. Merakli. "Risk-Averse Optimization in Multicriteria and Multistage Decision Making". PhD thesis. University of Washington, 2018.

[50] A. Modi, N. Jiang, S. Singh, and A. Tewari. "Markov decision processes with continuous side information". In: *Algorithmic Learning Theory*. 2018, pp. 597–618.

[51] Mount Hood 4 Modeling Group. "Computer Modeling of Diabetes and Its Complications". In: *Diabetes Care* 30.6 (2007).

[52] A. Nilim and L. El Ghaoui. "Robust Control of Markov Decision Processes with Uncertain Transition Matrices". In: *Operations Research* 53.5 (2005), pp. 780–798.

[53] L. Ntaimo. "Disjunctive Decomposition for Two-Stage Stochastic Mixed-Binary Programs with Random Recourse". In: *Operations Research* (2010). ISSN: 0030-364X. DOI: 10.1287/opre.1090.0693.

[54] C. H. Papadimitriou and J. N. Tsitsiklis. "The complexity of Markov decision processes". In: *Mathematics of Operations Research* 12.3 (1987), pp. 441–450.

[55] M. Pignone, S. Earnshaw, J. A. Tice, and M. J. Pletcher. "Aspirin, statins, or both drugs for the primary prevention of coronary heart disease events in men: a cost–utility analysis". In: *Annals of Internal Medicine* 144.5 (2006), pp. 326–336.

[56] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1994, p. 672.

[57] K. L. Rascati. "The $64,000 questionWhat is a quality-adjusted life-year worth?" In: *Clinical Therapeutics* 28.7 (2006), pp. 1042–1043. ISSN: 0149-2918. DOI: https://doi.org/10.1016/j.clinthera.2006.07.002. URL: http://www.sciencedirect.com/science/article/pii/S0149291806001640.

[58] B. Roy. "Robustness in operational research and decision aiding: A multi-faceted issue". In: *European Journal of Operational Research* 200.3 (2010), pp. 629–638. ISSN: 03772217. DOI: 10.1016/j.ejor.2008.12.036. URL: http://dx.doi.org/10.1016/j.ejor.2008.12.036.

[59] S. Saghafian. "Ambiguous Partially Observable Markov Decision Processes: Structural Results and Applications". In: *Journal of Economic Theory* 178 (2018), pp. 1–35.

[60] J. K. Satia and R. E. Lave Jr. "Markovian decision processes with uncertain transition probabilities". In: *Operations Research* 21.3 (1973), pp. 728–740.

[61] L. J. Savage. *The foundations of statistics*. Courier Corporation, 1972.

[62] D. Scheftelowitsch, P. Buchholz, V. Hashemi, and H. Hermanns. "Multi-objective approaches to Markov decision processes with uncertain transition parameters". In: *arXiv preprint arXiv:1710.08986* (2017).

[63] A. Shapiro, D. Dentcheva, and A. P. Ruszczyski. *Lectures on stochastic programming : modeling and theory*. Society for Industrial and Applied Mathematics. Mathematical Optimization Society., 2009, p. 494. ISBN: 1611973422.

[64] S. M. Shechter, M. D. Bailey, a. J. Schaefer, and M. S. Roberts. "The Optimal Time to Initiate HIV Therapy Under Ordered Health States". In: *Operations Research* 56 (2008), pp. 20–33.

[65] S. P. Singh, T. Jaakkola, and M. I. Jordan. "Learning without state-estimation in partially observable Markovian decision processes". In: *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 284–292.

[66] S. Sinha, J. Kotas, and A. Ghate. "Robust response-guided dosing". In: *Operations Research Letters* 44.3 (2016), pp. 394–399.

[67] R. D. Smallwood and E. J. Sondik. "The Optimal Control of Partially Observable Markov Decision Processes over a Finite Horizon". In: *Operations Research* 21.5 (1973), pp. 1071–1088.

[68] V. Snow, M. D. Aronson, E. R. Hornbake, C. Mottur-Pilson, and K. B. Weiss. "Lipid control in the management of type 2 diabetes mellitus: a clinical practice guideline from the American College of Physicians". In: *Annals of Internal Medicine* 140.8 (2004), pp. 644–649.

[69] L. N. Steimle and B. T. Denton. *Markov decision processes for screening and treatment of chronic diseases*. Vol. 248. 2017, pp. 189–222.

[70] L. N. Steimle, D. L. Kaufman, and B. T. Denton. "Multi-model Markov Decision Processes". Optimization online. 2018.

[71] R. M. Van Slyke and R. Wets. "L -Shaped Linear Programs with Applications to Optimal Control and Stochastic Programming". In: *SIAM Journal on Applied Mathematics* (1969).

[72] N. Vlassis, M. L. Littman, and D. Barber. "On the Computational Complexity of Stochastic Controller Optimization in POMDPs". In: *ACM Transactions on Computation Theory* 4.4 (Nov. 2012), pp. 1–8.

[73] C. C. White III and H. K. Eldeib. "Markov decision processes with imprecise transition probabilities". In: *Operations Research* 42.4 (1994), pp. 739–749.

[74] W. Wiesemann, D. Kuhn, and B. Rustem. "Robust Markov decision processes". In: *Mathematics of Operations Research* 38.1 (2013), pp. 153–183.

[75] P. W. F. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel. "Prediction of Coronary Heart Disease Using Risk Factor Categories". In: *Circulation* 97.18 (1998), pp. 1837–1847.

[76] P. A. Wolf, R. B. D'Agostino, A. J. Belanger, and W. B. Kannel. "Probability of stroke: a risk profile from the Framingham Study." In: *Stroke* 22.3 (1991), pp. 312–318.

[77] R. D. Wollmer. "Two stage linear programming under uncertainty with 0-1 integer first stage variables". In: *Mathematical Programming* (1980). ISSN: 00255610. DOI: 10.1007/BF01581648.

[78] H. Xu, S. Mannor, et al. "Distributionally Robust Markov Decision Processes". In: *Mathematics of Operations Research* 37.2 (2012), pp. 288–300.

[79] H. Xu, S. Mannor, et al. "Parametric regret in uncertain markov decision processes". In: *In IEEE Conference on Decision and Control, CDC*. Citeseer. 2009.

[80] Y. Zhang, L. Steimle, and B. T. Denton. "Robust Markov Decision Processes for Medical Treatment Decisions". In: *Optimization online.* (2017).

[81] Y. Zhang, H. Wu, B. T. Denton, J. R. Wilson, and J. M. Lobo. "Probabilistic sensitivity analysis on Markov models with uncertain transition probabilities: an application in evaluating treatment decisions for type 2 diabetes". In: *Health Care Management Science* (2017), pp. 1–19.