IISE Transactions

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uiie21

Multi-model Markov decision processes

Lauren N. Steimle, David L. Kaufman & Brian T. Denton

To cite this article: Lauren N. Steimle, David L. Kaufman & Brian T. Denton (2021): Multi-model Markov decision processes, IISE Transactions, DOI: 10.1080/24725854.2021.1895454

To link to this article: https://doi.org/10.1080/24725854.2021.1895454

IISE TRANSACTIONS
https://doi.org/10.1080/24725854.2021.1895454

Multi-model Markov decision processes

Lauren N. Steimle[a], David L. Kaufman[b], and Brian T. Denton[c]

[a]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA; [b]College of Business, University of Michigan–Dearborn, Dearborn, MI, USA; [c]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Markov decision processes (MDPs) have 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 data used to parameterize the model can influence what policies are recommended, and multiple competing data sources are common in many application areas, including medicine. In this article, we introduce the Multi-model Markov decision process (MMDP) which generalizes a standard MDP by allowing for multiple models of the rewards and transition probabilities. Solution of the MMDP generates a single policy that maximizes the weighted performance over all models. This approach allows the decision maker to explicitly trade-off conflicting sources of data while generating a policy of the same level of complexity for models that only consider a single source of data. We study the structural properties of this problem and show that it is at least NP-hard. 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.

ARTICLE HISTORY
Received 10 January 2020
Accepted 24 January 2021

KEYWORDS
Dynamic programming; medical decision making; Markov decision processes; parameter ambiguity; healthcare applications

1. Introduction

The Markov decision processs (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 (Puterman, 2014; Boucherie and van Dijk, 2017). MDPs generalize Markov chains in that a decision maker (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 due to these values being 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 much worse than anticipated (Mannor *et al.*, 2007). This motivates the need for MDPs that account for this ambiguity in the MDP parameters.

In this article, 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 such a 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.

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 2010 to 2020 revealed more than 8300 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 (Alagoz *et al.*, 2007), HIV (Shechter *et al.*, 2008), breast cancer (Ayer *et al.*, 2012), and diabetes (Mason *et al.*, 2014).

Despite the potential of MDPs to inform medical decision making, the utility of these models is often at the

CONTACT Lauren N. Steimle steimle@gatech.edu
Supplemental data for this article is available online at https://doi.org/10.1080/24725854.2021.1895454.

mercy of the data available to parameterize the models. The transition dynamics in medical decision making models are commonly 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 (2007), Etzioni *et al.* (2012), and Mandelblatt *et al.* (2016) 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 cardiovascular disease for which 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 (Goff *et al.*, 2014) and the risk equations resulting from the Framingham Heart Study (FHS) (Wolf *et al.*, 1991; Wilson *et al.*, 1998). However, these two credible models give conflicting estimates of a patient's risk of having a major cardiovascular event. Steimle and Denton (2017) showed that the best treatment protocol for cardiovascular disease 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 *et al.* (2018)), 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 (Habbema *et al.*, 2006).

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 programming methods to account for parameter ambiguity in MDPs? Further, how much benefit is there to mitigating the effects of ambiguity?

1.2. Contributions

In this article, we present a new approach for handling parameter ambiguity in MDPs, which we refer to as the Multi-model Markov decision process (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 a common action space. We consider a problem in which each model has a corresponding weight, and the DM seeks to find a single policy that will maximize the weighted value function.

We show that, in general, optimal policies that maximize the weighted value function may actually be history dependent, making the problem of maximizing the weighted value function more challenging to solve in certain cases. With the aim of designing policies that are easily translated to practice, we distinguish between two important variants: (i) a case where the DM is limited to policies determined by the current state of the system, which we refer to as the *Weighted Value Problem (WVP)*, and (ii) 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* counterpart to the WVP. We show that the adaptive counterpart is a special case of a Partially-Observable MDP (POMDP) that is PSPACE-hard, and we show that the WVP is NP-hard.

Based on our complexity analysis, the well-known backward induction algorithm for finite-horizon MDPs cannot solve the WVP to optimality. Therefore, we 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. For larger problem instances, as one might find in medical decision making applications, models are computationally intractable. Therefore, we introduce a fast heuristic based on backwards recursion that we refer to as the Weight-Select-Update (WSU) with computational bounds on the error. The WSU heuristic is fast and scales to larger medical decision making instances, such as the instance that motivated this work.

Finally, we present a case study for prevention of cardiovascular disease, a setting in which there is ambiguity due to the existence of two well-known and competing risk models for cardiovascular events (ACC/AHA and FHS). The goal is to design an optimal treatment guideline that would work well from a population perspective given both models are plausibly correct. We show that this problem can me modeled as an MMDP and solve the corresponding WVP. 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 being robust to ambiguity arising from the existence of multiple conflicting models.

The remainder of this article is organized as follows: In Section 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 3, and in Section 4 we present analysis of our proposed WVP for the MMDP model. In Section 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 6. In Section 7, we present our case study. Finally, in Section 8, we summarize the most important findings from our research and discuss the limitations and opportunities for future research.

2. Background and literature review

In this article, we focus on discrete-time, finite-horizon MDPs with parameter 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.

2.1. MDPs

The MDP is a common framework for modeling sequential decision making that influences a stochastic reward process. The sequence of events that define the MDP are as follows: first, an initial state of the system $s_1 \in \mathcal{S}$ is determined 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 a new state of the system $s_2 \in \mathcal{S}$ is realized with probability $p_1(s_2|s_1, a_1) \in [0, 1]$. This process continues whereby for any *decision epoch* $t \in \mathcal{T} \equiv \{1, ..., T\}$, 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 a new state $s_{t+1} \in \mathcal{S}$ is realized with probability $p_t(s_{t+1}|s_t, a_t)$. The DM selects the last action at time T that 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 article, we assume, without loss of generality, that the discount factor is already incorporated into the reward definition. We will refer to the set of decision epochs as $\mathcal{T}$, 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 $p_t(s_{t+1}|s_t, a_t) \in [0, 1]$ and $\sum_{s_{t+1}\in\mathcal{S}} p_t(s_{t+1}|s_t, a_t) = 1, \forall t \in \mathcal{T}, s_t \in \mathcal{S}, a_t \in \mathcal{A}$. Throughout the remainder of this article, 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:

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

The objective of the DM is to select the sequence of actions in a strategic way so that the expectation of (1) 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 from which a DM may select: (i) history-dependent vs. Markov, and (ii) 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 $t \in \mathcal{T}$, whereas 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. For standard MDPs, there is guaranteed to be a Markov deterministic policy that maximizes the expectation of (1) (Proposition 4.4.3 of Puterman (2014)), 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:

$$\max_{\pi\in\Pi} \mathbb{E}^{\pi, P, \mu_1}\left[\sum_{t=1}^{T} r_t(s_t, a_t) + r_{T+1}(s_{T+1})\right]. \tag{2}$$

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. 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 1, the model parameters themselves may not be known with certainty. For clarity, throughout this article, 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 article, we consider a variation of MDPs 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 model corresponds to a "scenario" representing a different possibility for the problem data (Birge and Louveaux, 2011). 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 the MMDP, the DM must specify all actions before the model parameters are realized.

A recent stream of research on MDPs with parameter ambiguity has taken the approach of multiple models. Ahmed *et al.* (2017) proposed sampling rewards and transition probabilities at each time step to generate a finite set of MDPs and then seek to find one policy that minimizes the maximum regret over the set of MDPs. To do this, they formulate an 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 rectangularity property that is often leveraged in *robust dynamic programming* approaches (see Appendix A.1). Concurrently and independent of our work, Buchholz and Scheftelowitsch (2019) considered the problem of finding a policy that maximizes a weighted performance across "concurrent" infinite-horizon MDPs. They show that their problem is NP-hard and that randomized policies may be optimal in the infinite-horizon case. We will show that the finite-horizon problem is NP-hard and that there will exist a deterministic policy that is optimal. Building on the weighted value problem proposed here and by Buchholz and Scheftelowitsch (2019), Meraklı and Küçükyavuz (2020) proposed a percentile optimization formulation of the multiple models problem to reflect the DM with an aversion to losses in performance due to parameter ambiguity in infinite-horizon MDPs and Steimle *et al.* (2021) studied computational methods for solving our non-adaptive problem exactly. Meraklı and Küçükyavuz (2020) and Buchholz and Scheftelowitsch (2019) both provide mixed-integer linear programming formulations for determining the optimal pure policy and a nonlinear programming formulation for the optimal randomized policy, as well as local search heuristics that work well on their benchmark test instances. Multiple models have also been studied for POMDPs: Saghafian (2018) uses multiple models of the parameters to address ambiguity in transitions among the core states in a POMDP and uses an objective function that weights the best-case and worst-case value-to-go across the models. This is in contrast with our work that considers the expected value-to-go among multiple models. They assume that the best-case and worst-case model are selected independently across decision epochs. In our proposed MMDP formulation, the objective is to find a single policy that will perform well in each of the models which may have interdependent transition probabilities across different states, actions, and decision epochs.

Perhaps the most closely related healthcare-focused research to this article is that of Bertsimas *et al.* (2018) 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 (2002) 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, especially with the goal of these policies being just as easily translated to practice as those found by optimizing with respect to a single model.

In light of the medical decision application we discuss later, it should be noted that several past studies have considered aspects of model uncertainty in the healthcare context. Further, there have been other studies that have investigated multiple models in the area of *multi-task reinforcement learning*, and *robust Markov decision processes* are another approach for solving MDPs with parameter ambiguity. In the interest of brevity, we summarize this literature in Appendix A. Although our approach is distinct from the robust MDP approaches described in the appendix, we provide a comparison of how these two approaches compare on our case study (and the details are in Appendix E).

3. Multi-model Markov decision processes

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

Definition 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 $\Lambda := \{\lambda_1, ..., \lambda_{|\mathcal{M}|}\}$ 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, are exogenous, and it is assumed that the DM has accurate estimates of these weights. Depending on the context, the model weights may be determined by expert judgment, estimated from empirical distributions, or treated as uninformed priors when each model is considered equally reputable (as in Bertsimas *et al.* (2018)).

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:

$$v^m(\pi) := \mathbb{E}^{\pi, P^m, \mu_1^m}\left[\sum_{t=1}^{T} r_t^m(s_t, a_t) + r_{T+1}^m(s_{T+1})\right]. \quad (3)$$

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

$$W(\pi) := \sum_{m\in\mathcal{M}} \lambda_m v^m(\pi) = \sum_{m\in\mathcal{M}} \lambda_m \mathbb{E}^{\pi, P^m, \mu_1^m}\left[\sum_{t=1}^{T} r_t^m(s_t, a_t) + r_{T+1}^m(s_{T+1})\right]. \quad (4)$$

Thus, we consider the WVP in which the goal of the DM is to find the policy $\pi \in \Pi$ that maximizes the weighted value defined in (4):

Definition 2 *(Weighted Value Problem)*

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

$$W^* := \max_{\pi\in\Pi} W(\pi) = \max_{\pi\in\Pi}\left\{\sum_{m\in\mathcal{M}} \lambda_m \mathbb{E}^{\pi, P^m, \mu_1^m}\left[\sum_{t=1}^{T} r_t^m(s_t, a_t) + r_{T+1}^m(s_{T+1})\right]\right\} \quad (5)$$

and a set of policies $\Pi^* := \{\pi^* : W(\pi^*) = W^*\} \subseteq \Pi$ that achieve the maximum in (5).

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 that 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 of the system, and nature selects which model will be given to the DM by randomly sampling according to the model weights $\Lambda \in \mathscr{M}(\mathcal{M})$. In this sense, we might associate the model weights with a probability distribution over the models. 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 (5) when $\Pi = \Pi^M$. Here, the DM specifies a Markov policy, $\pi \in \Pi^M$, *a priori*. That is, 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: $\pi = \{\pi_t(s_t) = (\pi_t(1|s_t), \ldots, \pi_t(|\mathcal{A}||s_t)) \in \mathscr{M}(\mathcal{A}) : a_t \in \mathcal{A}, s_t \in \mathcal{S}, t \in \mathcal{T}\}$. In this policy, $\pi_t(a_t|s_t)$ 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. The choice of model remains fixed for the entire horizon. Now, $s_1 \in \mathcal{S}$ is determined 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, the next state $s_2 \in \mathcal{S}$ is determined according to $p_1^m(\cdot|s_1, a_1) \in \mathscr{M}(\mathcal{S})$. The interaction carries on in this way

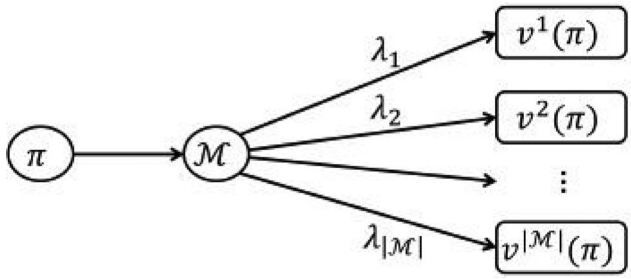

Figure 1. The figure illustrates the evaluation of a policy in terms of weighted value, which is the objective function used to compare policies for an MMDP. The DM specifies a policy π that is subsequently evaluated in each of the $|\mathcal{M}|$ models. The weighted value of a policy π is determined 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.

where the DM selects actions according to the pre-specified policy, π, and the next state is determined 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.

The evaluation of a given policy in the weighted value problem is illustrated in Figure 1. Policy evaluation is straightforward; one can use backwards recursion. Although policy evaluation is similar for MMDPs as compared to standard MDPs, policy optimization is much more challenging for the WVP. For example, backwards induction, a well-known solution technique for finite-horizon MDPs, does not apply to MMDPs where actions are coupled across models.

As mentioned above, in this article we focus on the WVP wherein the DM selects a Markov policy which can be interpreted as a non-adaptive problem. However, in general, the DM may benefit from a history-dependent policy that arises in the adaptive counterpart to the WVP. Although this is not the focus of this article, we consider this extension in which the DM considers all history-dependent policies in Appendix B. Some of the most important properties include that a deterministic optimal policy exists and that this problem is a special case of a POMDP. These findings allow us to reformulate the MMDP as an MDP defined on a continuous state space for which a Markov policy is optimal and design a solution method that leverages the special structure of the state space.

4. Analysis of MMDPs

In this section, we will analyze the WVP as defined in (5). 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 1. For ease of reading, we defer all proofs to Appendix C.

We begin by establishing the important result that there always exists a deterministic optimal policy for the WVP. 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

Table 1. Summary of the main properties and solution methods related to the MMDP.

Property	*Result*	*Support for Result*
Always a deterministic policy that is optimal?	Yes	Proposition 1
Computational complexity	NP-hard	Proposition 2
Exact solution method	MIP	Proposition 3
Upper bound on optimal weighted value?	Yes	Proposition 4
Heuristic, WSU		Procedure 1
Guaranteed to be optimal?	No	Proposition 5
Bound on the error when $\|\mathcal{M}\| = 2$?	Yes	Proposition 6

policy in this class and it could be desirable to find a policy with the same properties as 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 DM. 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 7, deterministic policies are a necessity, since randomization is unlikely to be considered ethical outside the context of randomized clinical trials.

Proposition 1. *There is always a Markov deterministic policy that is optimal for the WVP.*

This result means that for the WVP, 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 Singh *et al.* (1994), which states that the best stationary randomized policy can be arbitrarily better than the best stationary deterministic policy for POMDPs. Although this fact may seem to contradict Proposition 1, it is worth noting that Fact 2 of Singh *et al.* (1994) 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, no state s_t can be visited more than once.

Although policy evaluation is easy for the WVP, policy optimization over the class of Markov policies is provably hard.

Proposition 2. *Solving the WVP is NP-hard.*

We note that we have developed this result independently of a proof of an equivalent result which can be found in the thesis of Le Tallec (2007) describing the complexity of MDPs with "general random uncertainty". The proof in Appendix C provides a thorough description of the reduction required to prove this result. The result of Proposition 2 implies that we cannot expect to find an algorithm that solves the WVP for all MMDPs in polynomial time. Still, we are able to solve the WVP by formulating it as an MIP as discussed in the following proposition.

Proposition 3. *The WVP can be formulated as the following MIP:*

$$\max_{\pi, v} \sum_{m\in\mathcal{M}} \sum_{s\in\mathcal{S}} \lambda_m \mu_1^m(s) v_1^m(s) \tag{6a}$$

$$\text{s.t.} \quad \sum_{a\in\mathcal{A}} \pi_t(a|s) = 1, \quad \forall s \in \mathcal{S}, t \in \mathcal{T} \tag{6b}$$

$$M\pi_t(a|s) + v_t^m(s) - \sum_{s'\in\mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \le r_t^m(s,a) + M \tag{6c}$$

$$\forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T}$$
$$v_{T+1}^m(s) \le r_{T+1}^m(s), \quad \forall m \in \mathcal{M}, s \in \mathcal{S} \tag{6d}$$

$$\pi_t(a|s) \in \{0,1\}, \quad \forall a \in \mathcal{A}, s \in \mathcal{S}, t \in \mathcal{T} \tag{6e}$$

$$v_t^m(s) \text{ unrestricted}, \quad \forall s \in \mathcal{S}, t \in \mathcal{T}, m \in \mathcal{M}. \tag{6f}$$

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

It is well-known that standard MDPs can be solved using a linear programming (LP) formulation (Puterman, 2014, section 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 $v_t(s) \ge v_t(s,a)$ for all feasible a. In this reformulation, the constraint $v_t(s) \ge v_t(s,a)$ is tight for all actions that are optimal. The MIP formulation presented in (6a) 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 WVP, we require that constraints:

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

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 $v_t^m(s) \le v_t^m(s,a)$. Therefore, we must select M sufficiently large enough for all constraints.

The formulation of the WVP as an MIP may seem more natural after a discussion of the connections with two-stage stochastic programming (Birge and Louveaux, 2011). If we view the WVP 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. From a stochastic programming point of view, we can define a second-stage value function for a given set of first-stage decision variables π and a given realization of the model, m:

$$\begin{aligned} V(\pi, m) \quad &= \max_{v} \Big[\sum_{s \in \mathcal{S}} \mu_1^m(s) v_1^m(s) | M \pi_t(a|s) + v_t^m(s) \\ &\quad - \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \le r_t^m(s,a) + M, \\ &\quad \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T} \Big]. \end{aligned}$$

We also can define $V(\pi)$ as the *recourse function*:

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

Formulation (6a) is the deterministic equivalent formulation of this stochastic integer program. Notice that the second-stage value function, $V(\pi, m)$, corresponds to the definition of $v^m(\pi)$ in (3), and the recource function, $V(\pi)$, corresponds to the weighted value of a policy $W(\pi)$ defined in (4). The view of the WVP as a stochastic program also promotes a heuristic in which the DM solves the Mean Value Problem (MVP): a single MDP in which each parameter takes on its weighted averages from the corresponding parameters of the different models in the MMDP.

Our initial numerical experiments showed that moderate-sized MDPs can be solved using (6), 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 a heuristic that we describe in Section 5. The following relaxation of the WVP allows us to quantify the performance of our heuristic:

Proposition 4. *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,*

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

The result of Proposition 4 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 7.

Proposition 4 motivates several connections between robustness and the value of information. First, the upper bound in Proposition 4 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:

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

Although the wait-and-see value provides an upper bound, the value corresponds to a set of solutions, one for each model, rather than a single 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 create a measure of robustness for an MMDP termed the *Value of the Weighted Value Solution* (VWV):

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

which parallels the well-known *Value of the Stochastic Solution* (VSS) in stochastic programming (Birge and Louveaux, 2011, section 4.2). If VWV is low, this implies that there is not much value from solving the MMDP versus the MVP. On the other hand, if VWV is high, this implies that the DM will benefit significantly from solving the MMDP.

Although the WVP for MMDPs has connections to stochastic programming, it also has connections to POMDPs that are described in Appendix B. The MMDP may be constructed as a special case of a POMDP, in which the core states are comprised of a copy of each state corresponding to each model. The WVP described in the main body of this article can be viewed as the problem of finding the best *memoryless controller* for this POMDP (Vlassis *et al.*, 2012). 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). As no history is incorporated into the definition of the policy, this policy is permissible for the WVP.

5. Solution methods

We now discuss how to leverage the results of Section 4 to solve the WVP. For conciseness, we defer the solution methods for the adaptive counterpart to the WVP to Appendix B. We discuss the MIP formulation of Proposition 3 for solving the WVP. Although the MIP formulation provides a viable way to exactly solve this class of problems, the result of Proposition 2 motivates the need for a fast approximation algorithm that can scale to large MMDPs.

5.1. MIP 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:

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

Recall that the decision variables of the form $v_t^m(s) \in R$ 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 x 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:

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

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

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

Rearranging, we obtain:

$$M_{m,s,t} \ge r_t^m(s,a) + \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') - r_t^m(s,a') - \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a') v_{t+1}^m(s'), \quad \forall a, a' \in \mathcal{A}. \tag{7}$$

A sufficient condition for (7) is the following:

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

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.

Procedure 1 Weight-Select-Update (WSU) approximation algorithm

Input: MMDP

Let $\hat{v}_{T+1}^m(s_{T+1}) = r_{T+1}^m(s_{T+1}), \forall m \in \mathcal{M}$

$t \leftarrow T$

while $t \ge 1$ **do**

for Every state $s_t \in \mathcal{S}$ **do**

$$\hat{\pi}_t(s_t) \leftarrow \underset{a_t \in \mathcal{A}}{\operatorname{argmax}} \left\{ \sum_{m \in \mathcal{M}} \lambda_m \left(r_t^m(s_t, a_t) + \sum_{s_{t+1} \in \mathcal{S}} p_t^m(s_{t+1}|s_t, a_t) \hat{v}_{t+1}^m(s_{t+1}) \right) \right\} \tag{8}$$

end for

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

$$\hat{v}_t^m(s_t) \leftarrow r_t^m(s_t, \hat{\pi}_t(s_t)) + \sum_{s_{t+1} \in \mathcal{S}} p_t^m(s_{t+1}|s_t, \hat{\pi}_t(s_t)) \hat{v}_{t+1}^m(s_{t+1}) \tag{9}$$

end for

$t \leftarrow t-1$

end while

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

5.2. WSU approximation algorithm

Next, we discuss our WSU algorithm, formalized in Procedure 1, 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 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 5 shows that ignoring this could lead to sub-optimal selection of actions as illustrated in the proof.

Proposition 5. *WSU is not guaranteed to produce an optimal solution to the WVP.*

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 (8) may be suboptimal, the update of the value-to-go in each

model in (9) 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.

Lemma 1. *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*

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

Lemma 1 guarantees that the policies generated using WSU will have values in model $m \in \mathcal{M}$ that are non-decreasing in 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. Lemma 1 is also useful for establishing the lower bound in the following proposition:

Proposition 6. *For any MMDP with $|\mathcal{M}| = 2$, the error of the policy generated via WSU, $\hat{\pi}$, is bounded so that*

$$W(\pi^*) - W(\hat{\pi}) \leq \lambda_1\left(v^1(\pi^1) - v^1(\pi^2)\right) + \lambda_2\left(v^2(\pi^2) - v^2(\pi^1)\right),$$

where π^m is the optimal policy for model m and $\pi^* \in \Pi^{MD}$ is the optimal policy for WVP.

Proposition 6 provides an upper bound on the error for the important special case of two models. Unfortunately, the performance guarantee in Proposition 6 does not extend to $|\mathcal{M}| > 2$. The proof relies on Lemma 1 and the property that, when $|\mathcal{M}| = 2, \lambda_1 = 1 - \lambda_2$ to summarize the difference between two weight vectors in terms of a single parameter that which will satisfy a complete ordering. For $|\mathcal{M}| > 2$, this model-wise complete ordering is no longer available. Fortunately, the WSU heuristic and the upper bound of Proposition 4 together provide computational lower and upper bounds, respectively.

6. Computational experiments

In this section, we describe computational experiments involving a set of test instances for comparing solution methods for WVP on the basis of run-time and quality of the solution. The first set of experiments were based on a series of random instances of MMDPs. In Appendix D, we consider a second set of experiments that were based on a small MDP for determining the most cost-effective HIV treatment policy. To compare the solution methods, we generated a solution for each instance using the WSU heuristic, MVP heuristic, and the MIP formulation. We will compare the weighted value policies obtained via the heuristics ($W_N(\hat{\pi})$) 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. WSU and MVP were implemented using Python 3.7. All MIPs were solved using Gurobi 8.1.1.

6.1. Test instances

We now describe the test instances used to compare the solution methods. To generate the random test instances, first, the number of states, actions, models, and decision epochs were defined. We used a base case problem size of four states, four actions, four decision epochs, and four models. We generated a set of test instances where one aspect of the problem description was varied at a time to analyze the impact of growth in computation time as a function of states, actions, models, and decision epochs, independently. Once the number of states, actions, models, and decision epochs were defined, the model parameters were randomly sampled for each instance of the fixed problem size. 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 Dirichlet distribution (Dir), which has a set of parameters defining a base measure and a parameter defining the concentration of the distribution. For each row, the base measure was determined by sampling a uniform $U(0, 1)$ for each possible transition: $\tilde{p}(s'|s,a) \sim U(0,1)$. 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:

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

The $p(s'|s,a)$ values were then used as the base measure for the Dirichlet distribution, and we varied the concentration parameter to control for the amount of variation among the models. Dirichlet distributions with the same base measure have the same mean value of the transition row, but higher values of the concentration parameter correspond to distributions with less variance. For each sample, we scaled by a factor of $\beta \propto \min_{s' \in \mathcal{S}} p(s'|s,a)$ for $\beta = 1$, 10, and 100:

$$(p^m(1|s,a), \ldots, p^m(|\mathcal{S}||s,a)) \sim \text{Dir}(\beta p(1|s,a), \ldots, \beta p(|\mathcal{S}||s,a)),$$
$$\forall s \in \mathcal{S}, a \in \mathcal{A}, m \in \mathcal{M}.$$

These experiments allow us to test the performance of the solution methods on many different kinds of MMDPs; however, these instances are not guaranteed to have structured transitions and rewards that one might expect in practice. Therefore, we also include the following test instances that have a structure commonly observed in MDPs for medical decision making.

6.2. Results

We now present the results of our computational experiments comparing solution methods for WVP for these test instances. Appendix D.1 presents the run-time of the three proposed solution methods: MVP, WSU heuristic, and the

Table 2. The effect of the concentration parameter, β, on the performance of the WSU, MVP, and MIP solution methods on random MMDP test instances for the basecase instance size.

Concentration Parameter	Solution Time (CPU Seconds)						Optimality Gap (%)				VWV (%)		VWSU (%)		
	MIP		WSU		MVP		WSU		MVP						
	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Min.	Max.
1	5.73	18.54	< 0.01	< 0.01	< 0.01	< 0.01	0.42	2.81	1.03	4.11	1.67	7.70	0.78	−1.49	5.26
10	5.19	12.48	< 0.01	< 0.01	< 0.01	< 0.01	0.55	2.12	1.32	10.92	1.63	10.19	0.50	−4.38	9.35
100	4.80	9.81	< 0.01	< 0.01	< 0.01	< 0.01	0.32	2.23	1.42	8.23	1.23	14.34	0.65	−0.64	9.81

Notes. Each algorithm was run for 30 instances for each value of the concentration parameter. The Value of the Weighted Value problem (VWV) and the value of using the WSU heuristic (VWSU) are also reported for each value of the concentration parameter.

exact MIP formulation. In summary, we find that the MVP and WSU were able to solve the random test instances quickly (under 0.1 CPU seconds for each instance) whereas the average time to solve the MIP noticeably increased as the size of the problem increases. The results suggest that heuristics are needed to approximate solutions for larger MMDPs, such as the one presented in the case study in Section 7. The MVP and WSU heuristic also performed well in terms of their average optimality gaps, although WSU provided a better optimality gap in 79.2% of the test instances. WSU had an average optimality gap of 0.53% and worst-case gap of 10.17%, whereas the MVP had an average optimality gap of 1.17% and worst-case gap of 12.80%. Table 2 shows the effect of the concentration parameter, β, on the computational time and optimality gap. It appears that the solution time for the MIP decreases as the concentration parameter increases, but there is no such pattern between the solution time for the WSU and MVP solution times. Furthermore, there does not appear to be a clear connection between the concentration parameter and the optimality gap of the heuristics. In Table 2, we also report the average and maximum VWV for each concentration parameter value. We observe that the average VWV is over 1% for each value of the concentration parameter and the maximum VWV was over 14% across all instances. These findings suggest that the solution to the MMDP could be a valuable alternative to the solution of the MVP. Furthermore, we report the value of using the heuristic WSU rather than the MVP solution. We report this values as the value of the WSU policy minus the value of the MVP policy normalized to the value of the MVP policy. This is indicated as "VWSU(%)" in the table. On average, WSU performs better than MVP, but we do see cases where MVP can perform up to 4.38% better. However, there are cases where WSU can perform up to 9.81% better than MVP. Fortunately, both of these solution methods are fast, and therefore, a DM could generate both policies and pick the one that performs better. WSU and MVP also solve the medical decision making instances quickly and perform quite well in terms of maximum optimality gap (see Appendix D.2).

7. Case study: Blood pressure and cholesterol management 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. We begin 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 (Centers for Disease Control and Prevention, 2011). 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 trade-off 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 Lipids (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 (Wolf *et al.*, 1991; Wilson *et al.*, 1998) and the ACC/AHA assessment of cardiovascular risk (Goff *et al.*, 2014).

7.1. MMDP formulation

The MDP formulation of Mason *et al.* (2014) was adapted to create an MMDP based on the FHS risk model (Wolf *et al.*, 1991; Wilson *et al.*, 1998) and the ACC/AHA risk model (Goff *et al.*, 2014). 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,

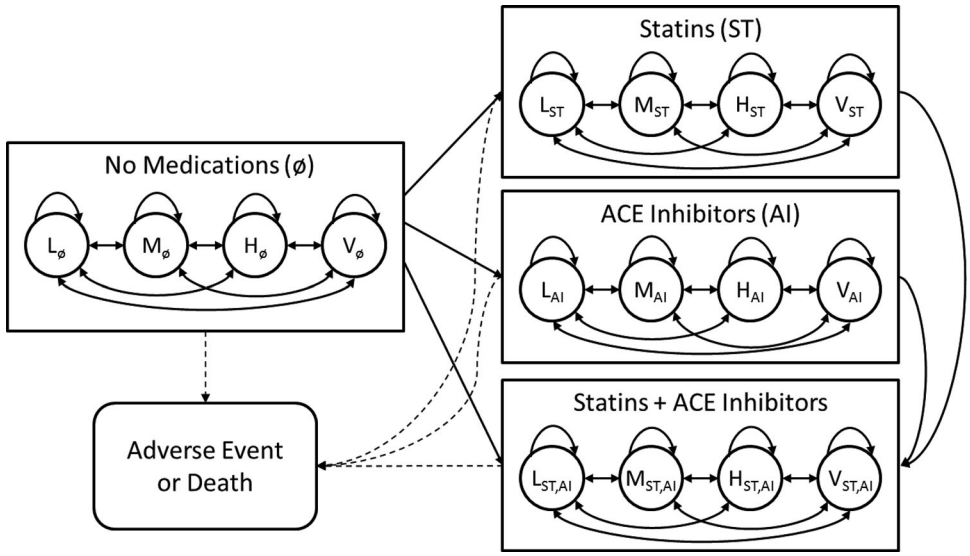

Figure 2. An illustration of the state and action spaces of the MDP as illustrated in Mason *et al.* (2014). In the corresponding MMDP, when medications are initiated (solid lines denote actions), the risk factors are improved and the probability of an adverse event (denoted by the dashed lines) is reduced. The probabilities of adverse events may differ in the different models depending on the risk calculator.

TC, HDL, and current medications. A discrete set of actions represents the initiation of the two cholesterol medications and four 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 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 $t \in \mathcal{T} = \{0, 1, ..., T\}$ that correspond to ages 54 through 74 at 1-year intervals that represent annual preventive care visits with a primary care doctor. These ages represent the median age of diagnosis of diabetes among patients in the calibrating dataset until the age for which the risk estimators provide predictions of cardiovascular risk. It is assumed that once a patient starts a medication, the patient will remain on this medication for the rest of his or her life, which is consistent with clinical recommendations (Chobanian *et al.*, 2003; Vijan and Hayward, 2004). 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 (Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, 2001). 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 $\mathcal{B} = \{\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_n) : \tau_i \in \{0, 1\}, \ \forall i = 1, 2, ..., 6\}$ corresponding to all combinations of the six 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 Mason *et al.* (2014). The living states in the model are indexed by $(\ell, \boldsymbol{\tau}) \in \mathcal{L} \times \mathcal{B}$. The absorbing states 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, 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 $\mathbf{A}_{(\ell,\boldsymbol{\tau})} = \{A_{(\ell,\tau_1)} \times A_{(\ell,\tau_2)} \times \cdots \times A_{(\ell,\tau_n)}\}$. Action $\mathbf{a} \in \mathbf{A}_{(\ell,\boldsymbol{\tau})}$ denotes the action in state $(\ell, \boldsymbol{\tau})$. If a patient is in living state $(\ell, \boldsymbol{\tau})$ and takes action $\mathbf{a}$, the new medication state is denoted by $\boldsymbol{\tau}'$, where τ_i' is set to one 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 $\omega^{\text{TC}}(i), \omega^{\text{HDL}}(i)$, and $\omega^{\text{SBP}}(i)$, 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

Table 3. Time to approximate a solution to the weighted problem using the WSU algorithm and to solve each of the nominal models using standard dynamic programming, in CPU seconds.

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

nonfatal). At epoch t, $\bar{p}_t^{\boldsymbol{\tau}}(d|\ell)$ denotes the probability of transition from state $(\ell, \boldsymbol{\tau}) \in \mathcal{L} \times \mathcal{B}$ 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 blood pressure and cholesterol adjusted for the absence of medication (Denton *et al.*, 2009). We define $p_t^{\boldsymbol{\tau}}(j|\ell)$ 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:

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

The two models of the MMDP represent the different cardiovascular risk calculators used to estimate the transition probabilities to the absorbing states: $\bar{p}_t^{\boldsymbol{\tau}}(d|i)$ 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 all other cause mortality from the Centers for Disease Control and Prevention life tables (Arias and Xu, 2011). The reward $r_t(\ell, \boldsymbol{\tau})$ for a patient in health state ℓ at epoch t is:

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

where $\mathcal{Q}(\ell, \boldsymbol{\tau}) = 1 - d^{\text{MED}}(\boldsymbol{\tau})$ is the reward for one QALY. QALYs are elicited through patient surveys, and are commonly used for health policy studies (Gold *et al.*, 2002). The *disutility* factor, $d^{\text{MED}}(\boldsymbol{\tau})$, represents the estimated decrease in quality of life due to the side effects associated with the medications in $\boldsymbol{\tau}$. We use the disutility estimates provided in Mason *et al.* (2014).

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. The MMDP had 4099 states, 64 actions, 20 decision epochs, and two models.

Table 3 shows the computation time required to run WSU with $\lambda_F = \lambda_A = 0.5$, as well as the time required to solve the FHS model and the ACC/AHA model using standard dynamic programming, for the female and male problem parameters. Although 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 3 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 4. When $\lambda_F = 100\%$, WSU finds the optimal policy for the FHS model, which is why the maximum of the FHS value is 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 the MVP. As supported by Proposition 1, 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

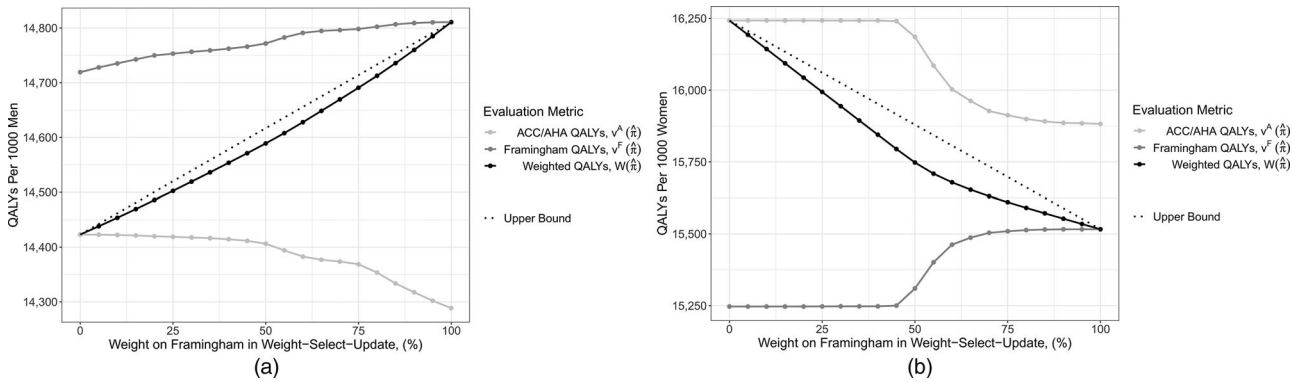

Figure 3. The performance of the policies generated using the WSU approximation algorithm for the MMDP for treatment of men (Figure 3(a)) and women (Figure 3(b)). For each choice of the weight on the FHS model in WSU, the graph shows the performance of these policies with respect to three different metrics: the performance in the ACC/AHA model (light gray), the performance in the FHS model (dark gray), and the weighted value (black). The dotted line represents the upper bound from Proposition 4.

Table 4. The performance of three policies in terms of QALYs gained over no treatment and regret.

(a) Male

Metric (per 1000 men)	*Evaluation*	*ACC/AHA Optimal Policy*	*FHS Optimal Policy*	*WSU Policy*
QALYs Gained	ACC/AHA	695.9	561.5	679.3
Over No Treatment	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	ACC/AHA	205.2	−155.3	147.9
Over No Treatment	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

Notes. The three policies are: (i) the optimal policy for the ACC/AHA model; (ii) the optimal policy for the FHS model; and (iii) the policy generated via the 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 policy for an individual model and the QALYs obtained by the given policy. The "Weighted" metrics report the value obtained by evaluating the given policy in each model separately, and then taking the weighted average.

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 3(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 Appendix A.1, 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 and 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 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: (i) the ACC/AHA model's optimal policy; (ii) the FHS model's optimal policy; and (iii) the WSU policy. Although 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*: the difference, for a specific model, 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 policy achieves 134.4 QALYs per 1000 men less than the ACC/AHA model's optimal policy, whereas 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 whereas 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 achieves 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. We discuss the policy associated with the solution generated using WSU when the weights are treated as an uninformed prior on the models in Appendix D.3. As discussed there, 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. Although 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. 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. In addition, the WSU algorithm generates a small gain over the policies found via the MVP approach. WSU led to a gain of 0.9 QALYs per 1000 women and a gain of 0.1 QALYs per 1000 men over the MVP policy.

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. As $WSU \leq W^*$, 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.

Although our case study was motivated by differences between the FHS and ACC/AHA risk models, our approach is easily generalized to more than two models. In Appendix E, we investigate the sensitivity of the results to the number of models in the MMDP by considering three different natural history scenarios in combination with the two cardiovascular risk models. The policies were quite different for models with different cardiovascular risk calculators, but differences in the natural history model did not play as large of a role. In this expanded MMDP model, the WSU policy performs better than each individual model in terms of weighted performance.

8. Conclusions

In this article, we addressed the following research questions: (i) how can we improve stochastic dynamic programming methods to account for parameter ambiguity in MDPs? (ii) how much benefit is there to mitigating the effects of ambiguity? To address the first question, we introduced the MMDP, which is an MDP with multiple models of the reward and transition probability parameters, and the WVP whose solution provides a policy that maximizes the weighted value across these models. We proved that the solution of the WVP 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 WVP for an MMDP 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 WVP.

To address the second research question, we established connections between concepts in stochastic programming and the WVP that quantify the impact of ambiguity on an MDP. We showed that the WVP 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 WVP as an MIP corresponding to the deterministic equivalent problem of the aforementioned two-stage stochastic program. We provided experiments comparing the results for the WVP considered in the main body and the adaptive counterpart considered in the appendix. We found the differences were small; however, our comparisons were based on small model instances. Whether or not there may be problems for which there are large differences between the WVP and its adaptive counterpart for large instances – which would be very difficult to solve efficiently due to the complexity of POMDPs – may be an opportunity for future research.

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 whereas 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 the WVP for an MMDP consisting of two models that 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 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.

Using our case study, we showed that the WSU generated small gains over the MVP. Although the difference is small for this particular case study, our numerical results in Section 6 showed the differences can be large in some cases. Although our results indicate WSU may be a safer choice, it is advisable to evaluate the WSU and MVP solutions, both of which have favorable computation time for large scale problems and significant benefits in terms of regret compared with selecting one of the MDP models arbitrarily. Future research could help identify conditions under which the MVP would generate a suitable solution versus conditions under which the MMDP approach could lead to extensive gains.

There are open opportunities for future work that builds off of the MMDP formulation. Future work could study the performance of the MMDP formulation for addressing statistical uncertainty compared with other robust formulations that have attempted to mitigate the effects of this kind of uncertainty. Another opportunity is to apply this approach to other diseases, such as diabetes, breast cancer, and prostate cancer, for which multiple models have been developed. Other future work might extend this concept to partially-observable MDPs and infinite-horizon MDPs, which are both commonly used for medical decision making. Furthermore, the bounds developed for the WSU were in the context of a two-model MMDP, but it could be valuable to develop bounds for WSU for $|\mathcal{M}| > 2$. Finally, the MMDP introduced in this article was limited to a finite number of models, however, future work may consider the possibility of a countably infinite number of models.

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.

Funding

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.

Notes on contributors

Lauren N. Steimle is an assistant professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology. She received her PhD and MSE in industrial and operations engineering from the University of Michigan and her BS in systems science and engineering from Washington University in St. Louis. Her research interests include data-driven optimization and stochastic modeling with applications to medicine and public health.

David Kaufman is an assistant professor at the University of Michigan-Dearborn, College of Business, where he teaches courses in Decision Sciences. He holds a PhD in Industrial and Operations Engineering from the University of Michigan. His research interests are in stochastic processes and decision models for systems where variability and uncertainty play an important role in design, analysis, and management.

Brian Denton is Chair of the Department of Industrial and Operations Engineering at the University of Michigan. His research interests are in data-driven sequential decision making and optimization under uncertainty with applications to medicine. Before joining the University of Michigan he worked at IBM, Mayo Clinic, and North Carolina State University. He is an INFORMS Fellow, past Chair of the INFORMS Health Applications Section, and he is Past President of INFORMS.

ORCID

Lauren N. Steimle http://orcid.org/0000-0002-4073-6165
David L. Kaufman http://orcid.org/0000-0002-0239-2920
Brian T. Denton http://orcid.org/0000-0002-6372-6066

References

Ahmed, A., Varakantham, P., Lowalekar, M., Adulyasak, Y. and Jaillet, P. (2017) Sampling based approaches for minimizing regret in uncertain Markov decision processes (MDPs). *Journal of Artificial Intelligence Research*, **59**, 229–264.

Alagoz, O., Maillart, L.M., Schaefer, A.J. and Roberts, M.S. (2007) Determining the acceptance of cadaveric livers using an implicit model of the waiting list. *Operations Research*, **55**(1), 24–36.

Arias, E. and Xu, J. (2011) National vital statistics reports United States life tables, 2007. *Statistics*, **59**(9), 1–132.

Ayer, T., Alagoz, O. and Stout, N.K. (2012) OR Forum - a POMDP approach to personalize mammography screening decisions. *Operations Research*, **60**(5), 1019–1034.

Bertsimas, D., Silberholz, J. and Trikalinos, T. (2018) Optimal healthcare decision making under multiple mathematical models: Application in prostate cancer screening. *Health Care Management Science*, **21**(1), 105–118.

Birge, J.R. and Louveaux, F. (2011) *Introduction to Stochastic Programming*, Springer, New York, NY.

Boucherie, R.J. and van Dijk, N.M. (eds) (2017) *Markov Decision Processes in Practice*. Cham, Switzerland: Springer.

Buchholz, P. and Scheftelowitsch, D. (2019) Computation of weighted sums of rewards for concurrent MDPs. *Mathematical Methods of Operations Research*, **89**(1), 1–42.

Centers for Disease Control and Prevention (2011) National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States. Technical report, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA.

Chobanian, A.V., Bakris, G.L., Black, H.R., Cushman, W.C., Green, L.A., Izzo, J.L., Jones, D.W., Materson, B.J., Oparil, S., Wright, J.T. and Roccella, E.J. (2003) Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*, **42**(6), 1206–1252.

Craig, B.A. and Sendi, P.P. (2002) Estimation of the transition matrix of a discrete-time Markov chain. *Health Economics*, **11**(1), 33–42.

Denton, B.T., Kurt, M., Shah, N.D., Bryant, S.C. and Smith, S.A. (2009) Optimizing the start time of statin therapy for patients with diabetes. *Medical Decision Making*, **29**(3),351–367.

Etzioni, R., Gulati, R., Tsodikov, A., Wever, E.M., Penson, D.F., Heijnsdijk, E.A., Katcher, J., Draisma, G., Feuer, E.J., de Koning, H.J. and Mariotto, A.B. (2012) The prostate cancer conundrum revisited: Treatment changes and prostate cancer mortality declines. *Cancer*, **118**(23), 5955–5963.

Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (2001) 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). *JAMA*, **285**(19), 2486–2497.

Goff, D.C., Lloyd-Jones, D.M., Bennett, G., Coady, S., D'Agostino, R.B., Gibbons, R., Greenland, P., Lackland, D.T., Levy, D., O'Donnell, C.J., Robinson, J.G., Schwartz, J.S., Shero, S.T., Smith, S.C., Sorlie, P., Stone, N.J. and Wilson, P.W.F. (2014) 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. *Circulation*, **129**, S49–S73.

Gold, M.R., Stevenson, D. and Fryback, D.G. (2002) HALYs and QALYs and DALYs, Oh My: Similarities and differences in summary measures of population health. *Annual Review of Public Health*, **23**(1), 115–134.

Habbema, J.D.F., Schechter, C.B., Cronin, K.A., Clarke, L.D. and Feuer, E.J. (2006) Chapter 16: Modeling cancer natural history, epidemiology, and control: Reflections on the CISNET Breast Group experience. *JNCI Monographs*, **2006**(36),122–126.

Le Tallec, Y. (2007) Robust, risk-sensitive, and data-driven control of Markov decision processes. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Mandelblatt, Jeanne, S., et al. (2016) Collaborative modeling of the benefits and harms associated with different US breast cancer screening strategies. *Annals of internal medicine*, **164**(4), 215–225.

Mannor, S., Simester, D., Peng, S. and Tsitsiklis, J.N. (2007) Bias and variance approximation in value function estimates. *Management Science*, **53**(2), 308–322.

Mason, J.E., Denton, B.T., Shah, N.D. and Smith, S.A. (2014) Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients. *European Journal of Operational Research*, **233**(3), 727–738.

Meraklı, M. and Küçükyavuz, S. (2020) Risk aversion to parameter uncertainty in Markov decision processes with an application to slow-onset disaster relief. *IISE Transactions*, **52**(8), 811–831.

Mount Hood 4 Modeling Group. (2007) Computer modeling of diabetes and its complications: a report on the Fourth Mount Hood Challenge Meeting. *Diabetes Care*, **30**(6), 1638–1646.

Puterman, M.L. (2014) *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Hoboken, NJ.

Saghafian, S. (2018) Ambiguous partially observable Markov decision processes: Structural results and applications. *Journal of Economic Theory*, **178**, 1–35.

Shechter, S.M., Bailey, M.D., Schaefer, A.J. and Roberts, M.S. (2008) The optimal time to initiate HIV therapy under ordered health states. *Operations Research*, **56**(1), 20–33.

Singh, S.P., Tommi, J. and Jordan, M.I. (1994) Learning without state-estimation in partially observable Markovian decision processes. *Proceedings of the Eleventh International Conference on Machine Learning 1994*, Morgan Kaufmann, 1994, pp. 284–292. San Francisco, CA, USA.

Steimle, L.N., Ahluwalia, V.S., Kamdar, C. and Denton, B.T. (2021) Decomposition methods for solving Markov decision processes with multiple models of the parameters. *IISE Transactions*, 1–37.

Steimle, L.N. and Denton, B.T. (2017) Markov decision processes for screening and treatment of chronic diseases, in *Markov Decision Processes in Practice*, Springer, New York, NY. https://doi.org/10.1080/24725854.2020.1869351

Vijan. S. and Hayward, R.A. (2004) Pharmacologic lipid-lowering therapy in type 2 diabetes mellitus: Background paper for the American College of Physicians. *Annals of Internal Medicine*, **140**(8), 650–658.

Vlassis, N., Littman, M.L. and Barber, D. (2012) On the computational complexity of stochastic controller optimization in POMDPs. *ACM Transactions on Computation Theory*, **4**(4), 1–8.

Wilson, P.W.F., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H. and Kannel, W.B. (1998) Prediction of coronary heart disease using risk factor categories. *Circulation*, **97**(18),1837–1847.

Wolf, P.A., D'Agostino, R.B., Belanger, A.J. and Kannel, W.B. (1991) Probability of stroke: A risk profile from the Framingham Study. *Stroke*, **22**(3), 312–318.