Appendix A: Other related literature

In this appendix, we provide a review of some additional literature that is related to our work of addressing parameter ambiguity in Markov decision processes.

A.1. Literature on parameter ambiguity in MDPs

The approach of incorporating multiple models of parameters is seen in the reinforcement learning literature, however the objective of the DM in these problems is different than the objective of the DM in this article. For example, consider what is perhaps the most closely related reinforcement learning problem: the *Contextual Markov Decision Process* (CMDP) proposed by Hallak et al. (2015). The CMDP is essentially the same as the MMDP set-up in that one can think of the CMDP as an integer number, C, of MDPs all defined on the same state space and action space, but with different reward and transition probability parameters. In the CMDP problem, the DM will interact with the CMDP throughout a series of episodes occurring serially in time. At the beginning of the interaction, the DM neither has any information about any of the C MDPs' parameters, nor does she know which MDP she is interacting with at the beginning of each episode. Our work differs from that of Hallak et al. (2015) in that we assume the DM has a complete characterization of each of the MDPs, but due to ambiguity the DM still does not know which MDP she is interacting with. Others have studied related problems in the setting of *multi-task reinforcement learning* (Brunskill and Li 2013). Our work differs from this line of research in that we are motivated by problems with shorter horizons while multi-task learning is appropriate for problems in which the planning horizon is sufficiently long to observe convergence of estimates to their true parameters based on a dynamic learning process.

Our work is also distinct from the more traditional approach of mitigating parameter ambiguity in MDPs known as *robust dynamic programming*. Iyengar (2005) and Nilim and El Ghaoui (2005) provide algorithms for solving the max-min problem by providing polynomial-time methods that assume that a rectangularity assumption is valid for the ambiguity set. While rectangular ambiguity sets are desirable from a computational perspective, they can give rise to policies that are overly-conservative because the DM must account for the possibility that parameters for each state-action-time triplet will take on their worst-case values simultaneously. Much of the research in robust dynamic programming has focused on ways to mitigate the effects of parameter ambiguity while avoiding policies that are overly conservative by either finding non-rectangular ambiguity sets that are tractable for the max-min problem or optimizing with respect to another objective function usually assuming some *a priori* information about the model parameters (Delage and Mannor 2009, Xu and Mannor 2012, Wiesemann et al. 2013, Mannor et al. 2016, Li et al. 2017, Scheftelowitsch et al. 2017, Goyal and Grand-Clement 2018).

A.2. Parameter ambiguity in medical decision making

To our knowledge, the optimal design of medical screening and treatment protocols under parameter ambiguity is limited to the work of Kaufman et al. (2011), Sinha et al. (2016), Zhang et al. (2017), and Boloori et al. (2020). Kaufman et al. (2011) consider the optimal timing of living-donor liver transplantations, for which some critical health state are seldom visited historically. They use the robust MDP framework, modeling ambiguity sets as confidence regions based on relative entropy bounds. The resulting robust solutions are of a simple control-limit form that suggest transplanting sooner, when patients are healthier, than otherwise suggested by traditional MDP solutions based on maximum likelihood estimates of transition probabilities. Sinha et al. (2016) use a robust MDP formulation for response-guided dosing decisions in which the dose-response parameter is allowed to vary within an interval uncertainty set and show that a monotone dosing policy is optimal for the robust MDP. Zhang et al. (2017) propose a robust MDP framework in which transition probabilities are confined to statistical confidence intervals. They employ a rectangularity assumption implying independence of rows in the transition probability matrix, and they assume an adversarial model in which the DM decides on a policy and an adversary optimizes the choice of transition probabilities that minimizes expected rewards subject to an uncertainty budget on the choice of transition probabilities. Boloori et al. (2020) leverages the results of Saghafian (2018) to inform decision-making related to immunosuppressive medication use for patients after organ transplantations to balance the risk of diabetes after transplantation and the risk of organ rejection. While these articles address parameter ambiguity in the transition probabilities, they all assume a rectangular ambiguity set which decouples the ambiguity across decision epochs and states. In contrast, the MMDP formulation that we propose allows a relaxation of this assumption. It allows for the ambiguity in model parameters to be linked across tuples of states, actions, and decision epochs.

Appendix B: Analysis of the adaptive problem

In this appendix, we present the adaptive counterpart of the WVP presented in the main body of this article. To distinguish, we refer to the adaptive counterpart of the WVP as the *adaptive problem* and the WVP as described in the main body as the *non-adaptive problem*. The main results related to the adaptive problem are presented in Table EC.1.

B.1. Problem statement

The adaptive problem generalizes the non-adaptive problem to allow the DM to utilize realizations of the states to adjust her strategy. In this problem, nature and the DM interact sequentially where the DM gets new information in each decision epoch of the MMDP and the DM is allowed to utilize the realizations of the states to infer information about the ambiguous problem parameters when selecting her future actions. In this setting, nature begins the interaction by selecting a model, $m \in \mathcal{M}$, according to the model weights Λ, and the model selected is not known to the DM. An initial state $s_1 \in \mathcal{S}$ is determined according to the model's initial distribution, μ_1^m. Next, the DM observes the state, s_1, and makes her move by selecting an action, $a_1 \in \mathcal{A}$. At this point, the next state, $s_2 \in \mathcal{S}$, is determined according to the distribution given by $p_1^m(\cdot|s_1, a_1) \in \mathscr{M}(\mathcal{S})$. The interaction continues where the DM observes the state and selects an action, and the next state is determined according to the distribution defined by the corresponding row of the transition probability matrix. For simplicity, we consider the adaptive problem only for the case where rewards are model-independent and transitions are model-dependent. We leave future analysis with model-dependent rewards to future research. It is important to note that the model weights are exogenous parameters, which are assumed to be known to the DM. That is, the DM precisely knows the probability distribution of nature. Other frameworks could instead assume some information asymmetry in which the DM did not know nature's distribution. We also leave this for future research.

B.2. Comparison of the adaptive and non-adaptive problems

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.

B.3. General properties of the weighted value problem

In both the adaptive and non-adaptive problems, nature is confined to the same set of rules. However, the set of strategies available to the DM in the non-adaptive problem is just a subset of the strategies available in the adaptive problem. Therefore, if W_N^* and W_A^* are the best expected values that the DM can achieve in the non-adaptive and adaptive problems, respectively, then it follows that $W_N^* \leq W_A^*$, and moreover, the inequality may be strict.

PROPOSITION EC.1. *It is possible that there are no optimal policies that are Markovian for the adaptive problem.*

The result of Proposition EC.1 is that the DM may benefit from being able to recall the history of the MMDP. This history allows for the DM to infer which model is most likely, conditional on the observed sample path, and tailor the future actions to reflect this changing belief about nature's choice of model. Therefore, the DM must search for policies within the history-dependent policy class to find an optimal solution to the adaptive MMDP. Hence, the adaptive problem does not reduce to the non-adaptive problem in general.

B.4. Analysis of the adaptive problem

We have shown that, in a similar way that POMDPs may have history-dependent policies that are optimal when they are defined on a discrete state space, the MMDP may also have a history-dependent policy in general when defined on the discrete state space. Just like POMDPS, we can reformulate the MMDP in such a way that would have a Markovian optimal policy, but this requires using a continuous-state representation. We begin by establishing an important connection between the adaptive problem and the POMDP (Smallwood and Sondik 1973):

PROPOSITION EC.2. *Any MMDP can be recast as a special case of a POMDP such that the maximum weighted value of the MMDP is equivalent to the expected discounted rewards of the POMDP.*

COROLLARY EC.1. *There is always a deterministic policy that is optimal for the adaptive problem.*

The implication of Proposition EC.2 is illustrated in Figure EC.1 which displays the relationship between MDPs, MMDPs, and POMDPs. Given Proposition EC.2, we can draw on similar ideas proposed in the literature for solving POMDPs and refine them to take advantage of structural properties specific to MMDPs. However, we show that even though MMDPs have special structure on the observation matrix and transition probability matrix (see the proof of Proposition EC.2 in

Table EC.1 Summary of the main properties and solution methods related to the adaptive WVP for the MMDP.

Property	**Result**	**Support**
Always a Markov policy that is optimal?	No	Proposition EC.1
Always a deterministic policy that is optimal?	Yes	Corollary EC.1
Computational Complexity	PSPACE-hard	Proposition EC.3
Exact solution method	Outer linearization	Procedure 2
	with state-wise pruning	Procedure 3

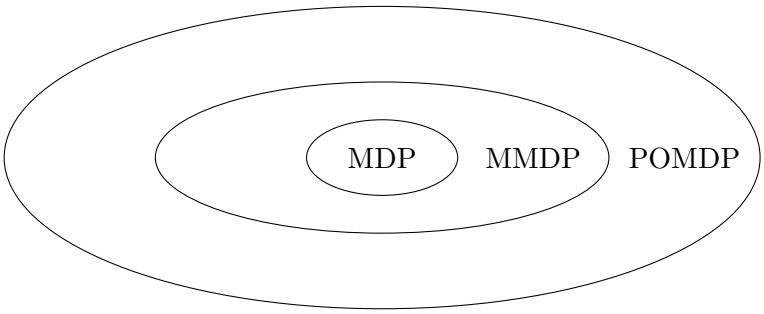


Figure EC.1 A Venn diagram illustrating the relationship between an MDP, MMDP, and POMDP. As shown in Proposition EC.2, any MMDP is a special case of a POMDP due to the structure of the transition matrix and observation conditional probabilities. Further, an MDP is a special case of an MMDP in which the MMDP only has one model.

Appendix C), we cannot expect any improvements in the complexity of the problem due to this structure. We note we have developed this proof independently of a proof of an equivalent result which was found in the thesis of Le Tallec (2007) describing the complexity of MDPs with "general random uncertainty".

PROPOSITION EC.3. *The adaptive problem for MMDPs is PSPACE-hard.*

Although the adaptive problem is PSPACE-hard and we cannot expect to develop an algorithm whose solution time is bounded above by a function that is polynomial in the problem size, we now discuss some special properties of the problem that can be exploited to develop an exact algorithm for solving this problem in Section 5. We start by establishing a sufficient statistic for MMDPs:

DEFINITION EC.1 (INFORMATION STATE FOR MMDPS). The information state for an MMDP is given by a vector:

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

with elements:

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

The fact that the information state is a sufficient statistic follows directly from Proposition EC.2, the formulation of a POMDP, and the special structure in the observation matrix.

Given this sufficient statistic, we establish some structural properties of the weighted value problem:

PROPOSITION EC.4. *The information state, b_t, has the following properties:*

1. *The value function is piece-wise linear and convex in the information state, b_t.*
2. $b_t(s,m) > 0 \Rightarrow b_t(s',m) = 0,\ \forall s' \neq s.$

3. *The information state as defined above is Markovian in that the information state b_{t+1} depends only on the information state and action at time t, b_t and a_t respectively, and the state observed at time t+1, s_{t+1}.*

According to part 1, the optimal value function can be expressed as the maximum value over a set of hyperplanes. This structural result forms the basis of our exact algorithm in Appendix B.5. Part 2 states that only elements in the vector with the same value for the state portion of the state-model pair (s,m) can be positive simultaneously, which implies that at most $|\mathcal{M}|$ elements of this vector are zero. This result allows us to ignore the parts of this continuous state space that have zero probability of being occupied. Part 3 allows for a sequential update of the belief that a given model is the best representation of the observed states given the DM's actions according to Bayes' rule. Consider the information state at time 1 at which point state s_1 has been observed. This information state can be represented by the vector with components:

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

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

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

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

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

As mentioned previously, our focus in this article is on applications of the MMDP framework to medical problems in contexts for which learning by Bayesian updating is not appropriate. However, the adaptive framework would apply to other contexts. We describe solution methods that exploit these structural properties in Appendix B.5.

B.5. Solution Methods for the Adaptive Problem

In this section, we describe an exact solution method that can be used to solve the adaptive problem for an MMDP. We begin by describing Procedure 2 which is an exact solution method for solving the adaptive weighted value problem. The correctness of this solution method follows from Proposition EC.2 which states that every MMDP is a special case of a POMDP and that the

maximum weighted value is equivalent to the expected discounted rewards of the corresponding POMDP. Therefore, we transform the MMDP into a POMDP and use a solution method analogous to a well-known solution method for POMDPs (Smallwood and Sondik 1973). This method exploits the property that the value function is piece-wise linear convex and therefore can be represented as the maximum over a set of supporting hyperplanes (Proposition EC.4).

In the worst case, the number of hyperplanes needed to represent the value function could potentially be as large as $1 + |\mathcal{A}| + \sum_{t=1}^{T-1} |\mathcal{A}|^{|\mathcal{S}|+T-t}$ for $T \geq 2$, but in many cases the number of hyperplanes that are actually needed to represent the optimal value function is much smaller. *Pruning* describes the methods by which hyperplanes that are not needed to represent the optimal value function are discarded. The pruning method described in Procedure 3 is based on the LP method described in Smallwood and Sondik (1973), but exploits the result of Proposition 2 for computational gain. This result states that only certain parts of the information space are reachable due to the special structure of the MMDP and this allows for the LP problems for pruning to be decomposed into a set of smaller LPs.

For this procedure, we will use the information state as defined in Definition EC.1 and define the following notation:

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

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

For every action, we define the block diagonal matrix:

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

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

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

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

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

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

The space of all information states at time t is

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

Procedure 2 is a backwards induction algorithm which generates a set of hyperplanes at each decision epoch. Procedure 3 eliminates hyperplanes that are not necessary to represent the optimal value function. The DM selects the optimal sequence of actions for the observed history in an analogous way to a POMDP: update the information state based on the observation and select the action corresponding to the maximizing hyperplane at this particular information state.

Procedure 2 Algorithm for solving the adaptive weighted value problem (2)

Input: MMDP

Initialize $\mathcal{B}_{T+1} = \{r_{T+1}\}$

The value-to-go at time $T+1$. $v_{T+1}(b_{T+1}) = \beta'_{T+1} b_{T+1},\, \forall b_{T+1} \in B_{T+1}$

$t \leftarrow T$

while $t \geq 0$ **do**

for Every action a_t **do**

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

end for

$\mathcal{B}_t \leftarrow \cup_{a_t \in \mathcal{A}} \mathcal{B}_t(a_t)$

State-wise Prune$(\mathcal{B}_t)$

The value-to-go at time t is $v_t(b_t) = \max_{\beta_t \in \mathcal{B}_t} \beta'_t b_t, \quad \forall b_t \in B_t$

$t \leftarrow t-1$

end while

Output: Collection $B_0, \ldots, B_T$

Remark EC.1. While the non-adaptive problem has connections to stochastic programming, it also has connections to POMDPs as described above. The non-adaptive problem 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

Procedure 3 State-wise Prune

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

for Every vector $\beta \in \mathcal{B}$ **do**

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

Let $\mathcal{B}(s) = \{\beta_s : \beta_s(m) = \beta(s,m), \beta \in \mathcal{B}\}$

Solve the LP (EC.1)

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

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

end for

end for

Output: $\mathcal{B}$

is based only on the most recent observation of the state (recall that the DM gets no information about the model part of the state-model pair). Because no history is allowed to be incorporated into the definition of the policy, this policy is permissible for the non-adaptive problem. These connections between MMDPs and stochastic programs and POMDPs allow us to better understand the complexity and potential solution methods for finding the best solution to the non-adaptive problem.

B.6. Computational Experiments

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

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

To solve the adaptive version of these instances, Procedure 2 with was used with pruning Procedure EC.1. Procedure 2 was implemented using Python using SciPy's `linprog` package to solve the linear programs in the pruning for Procedure 3. Procedure 2 was terminated if $|\mathcal{B}| > 10,000$. The non-adaptive problem was solved exactly using the MIP formulation in (6). Each algorithm was implemented using Python 3.5.2. All MIPs were solved using AMPL Version 20150815 and CPLEX 12.6.1.

Our experiments investigated the difference between the non-adaptive and adaptive WVP solution on a set of random instances of MMDPs with 2 states, 2 actions, 2 models for 2 to 5 decision epochs. For each choice of decision epochs, 30 random instances were generated for a total of 120 random instances. For each instance, the non-adaptive problem was solved using the MIP formulation and the adaptive problem was solved using Procedure 2 with pruning. These experiments revealed a very small gap between the adaptive and non-adaptive solutions. For these instances, the average gap (calculated as $\frac{W_A^* - W_N^*}{W_N^*} \times 100\%$) was less than 0.1% and the worst-case gap was less than 3%.

To investigate the gap between the solutions of the non-adaptive and the adaptive problems for larger problem sizes, we compared the non-adaptive solution obtained via the MIP to the upper bound from Proposition 4. A base case problem of 4 states, 4 actions, 4 models, and 4 decision epochs was defined. A variety of problem sizes were tested by changing one aspect of the base case problem size at a time. The number of states was varied from 4 to 10, the actions from 4 to 10, and the models from 4 to 10, for a total of 28 different problem sizes. For each problem size, 100 instances were generated for a total of 2800 random instances. Over these 2,800 random instances, the worst-case gap between the MIP solution and the upper bound was 5.01%, and the average gap was 0.46%. Furthermore, the upper bound from Proposition 4 can be used to bound the gap between the non-adaptive solution and the adaptive solution.

In summary, the results of our experiments show that on the small problems that we considered, the gap between the optimal adaptive solution and the optimal non-adaptive solution can be quite small. However, we hypothesize that there would be more value to solving the adaptive problem relative to the non-adaptive problem for longer time horizons, although the value would also depend on the problem characteristics.

Appendix C: Proofs

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

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

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

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

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

$$\max_{q_T(s_T)} \sum_{a_T \in \mathcal{A}} q_t(a_T|s_T) \sum_{m \in \mathcal{M}} \mu_T^\pi(s_T, m) \left[r_T^m(s_T, a_T) + \sum_{s_{T+1}} p^m(s_{T+1}|s_T, a_T) r_{T+1}^m(s_{T+1}) \right]$$
$$\text{s.t. } q_T(a_T|s_T) \geq 0,$$
$$\sum_{a_T \in \mathcal{A}} q_T(a_T|s_T) = 1,$$

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

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

$$\sum_{s_t \in \mathcal{S}} \max_q \sum_{a_t \in \mathcal{A}} q_t(a_t|s_t) \sum_{m \in \mathcal{M}} \mu_t^\pi(s_t, m) \left[r_t^m(s_t, a_t) + \sum_{s_{t+1}} p^m(s_{t+1}|s_t, a_t) v_{t+1}^m(s_{t+1}) \right]$$
$$\text{s.t. } q_t(a_t|s_t) \geq 0,$$
$$\sum_{a_t \in \mathcal{A}} q_t(a_t|s_t) = 1.$$

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

$$\max_{q_t(a_t|s_t)=1} \sum_{a_t \in \mathcal{A}} q_t(a_t|s_t) \sum_{m \in \mathcal{M}} \mu_t^{\pi}(s_t, m) \left[r_t^m(s_t, a_t) + \sum_{s_{t+1}} p^m(s_{t+1}|s_t, a_t) v_{t+1}^m(s_{t+1}) \right]$$
$$\text{s.t. } q_t(a_t|s_t) \geq 0,$$
$$\sum_{a_t \in \mathcal{A}} q_t(a_t|s_t) = 1.$$

This is a linear program so there will exist an extreme point solution that is optimal. This extreme point solution corresponds to a deterministic decision rule for decision epoch t. $\square$

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

Proof of Proposition 2. We show that any 3-CNF-SAT problem can be transformed into the problem of determining if there exists a Markov deterministic policy for an MMDP such that the weighted value is greater than zero. Let's suppose we have a 3-CNF-SAT instance: a set of variables $U = \{u_1, u_2, \ldots, u_n\}$ and a formula $E = C_1 \wedge C_2 \ldots \wedge C_m$. We will construct an MMDP with one decision epoch from this instance of 3-CNF-SAT. In the only decision epoch, the state space consists of one state per variable, u_i, $i = 1, \ldots, n$. At the terminal stage, there are two states labeled "T" and "F". There are no immediate rewards for this problem. For every state u_i, there are two actions *true* or $false$. The terminal rewards correspond to a cost of 0 for reaching the terminal state "T" and a cost of 1 upon reaching the terminal state "F".

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

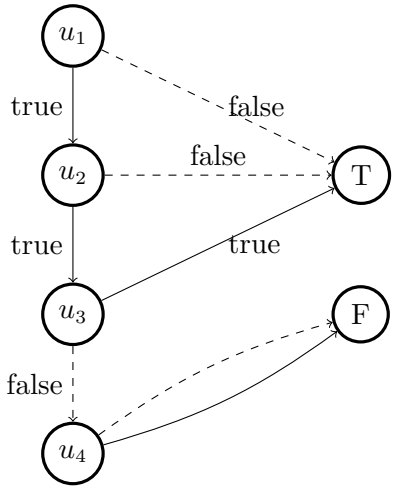
We will show that there is a truth assignment for the variables in U that satisfies E if and only if there is a Markov deterministic policy for the MMDP that achieves a weighted value equal to 0.

First, we show that if there is a truth assignment for the variables in U that satisfies E, then there exists a Markov deterministic policy for the MMDP that achieves a weighted value equal to 0. To construct such a policy, take the action *true* in every state u_i such that u_i is true is the satisfying truth assignment and take the action $false$ otherwise. Because this true assignment

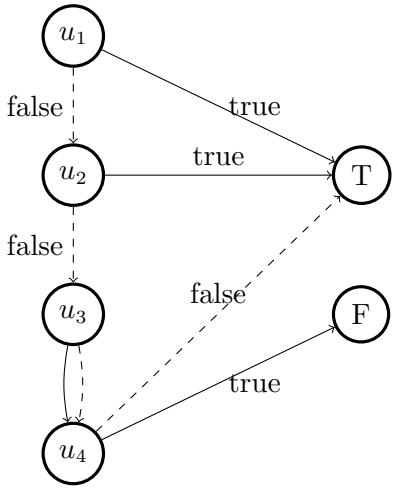
satisfies each clause, the corresponding policy will reach state "T" with probability 1 in each model. By construction, this policy will have a weighted value of zero.

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

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



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



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

Figure EC.2 An illustration of how a 3-CNF-SAT instance, $E = (u_1 \vee !u_2 \vee u_3) \wedge (u_1 \vee u_2 \vee !u_4)$, can be represented as an MMDP. Solid lines represent the transitions associated with the action $true$ and dashed lines represent the transitions associated with the action $false$. All transitions shown happen with probability 1.

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

$$\max_{\pi, v} \quad \sum_{m \in \mathcal{M}} \sum_{s \in \mathcal{S}} \lambda_m \mu_1^m(s) v_1^m(s) \tag{EC.2a}$$

$$\text{s.t.} \quad \sum_{a \in \mathcal{A}} \pi_t(a|s) = 1, \qquad \forall s \in \mathcal{S}, t \in \mathcal{T}, \tag{EC.2b}$$

$$M\pi_t(a|s) + v_t^m(s) - \sum_{s' \in \mathcal{S}} p_t^m(s'|s,a) v_{t+1}^m(s') \leq r_t^m(s,a) + M, \quad \text{(EC.2c)}$$
$$\forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T},$$
$$v_{T+1}^m(s) \leq r_{T+1}^m(s), \quad \forall m \in \mathcal{M}, s \in \mathcal{S}, \quad \text{(EC.2d)}$$
$$\pi_t(a|s) \in \{0,1\}, \quad \forall a \in \mathcal{A}, s \in \mathcal{S}, t \in \mathcal{T}, \quad \text{(EC.2e)}$$
$$v_t^m(s) \textit{ unrestricted}, \ \forall s \in \mathcal{S}, t \in \mathcal{T}, m \in \mathcal{M}. \quad \text{(EC.2f)}$$

Proof of Proposition 3 The decision variable $v_t(s)$ represents the optimal value-to-go for state $s \in \mathcal{S}$ at time $t \in \mathcal{T}$. The dual variables correspond to the probability of selecting an action given a state. Corner point solutions correspond to deterministic policies, and the optimal policy is deterministic by construction.

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

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,*

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

Proof of Proposition 4 The result follows from this series of inequalities:

$$\begin{aligned} \sum_{m \in \mathcal{M}} \lambda_m v^m(\hat{\pi}) &\leq \max_{\pi \in \Pi^{MD}} \sum_{m \in \mathcal{M}} \lambda_m v^m(\pi) \quad \text{(EC.3)} \\ &\leq \sum_{m \in \mathcal{M}} \lambda_m \max_{\pi \in \Pi^{MD}} v^m(\pi), \end{aligned}$$

where (EC.3) states than any MD policy will have a weighted value at most the optimal MD policy's weighted value. This optimal weighted value, in turn, is at most the value that can be achieved by solving each model separately and then weighting these values. □

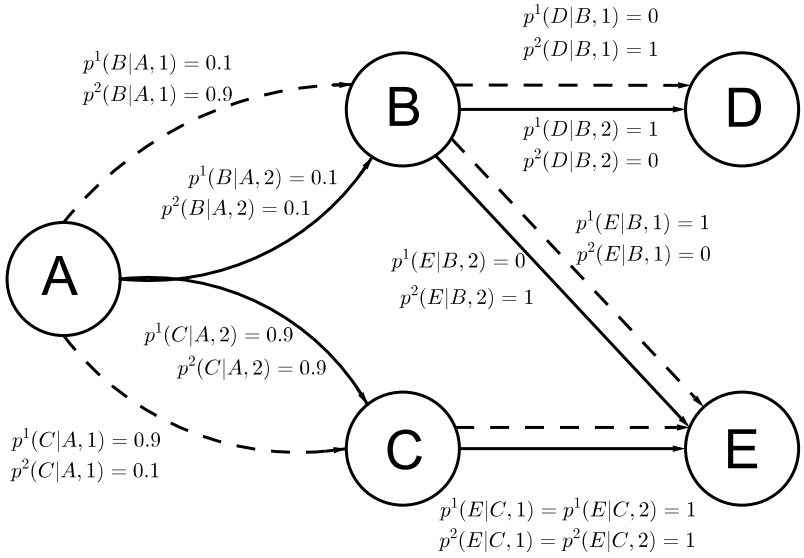


Figure EC.3 An illustration of an MMDP for which the WSU approximation algorithm does not generate an optimal solution to the non-adaptive weighted value problem. Possible transitions for actions 1 and 2 are illustrated with the dashed and solid line respectively. The probability of each possible transition in both of the models is listed by the corresponding line. The DM receives a reward of 1 if state D is reached. Otherwise, no rewards are received.

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

Proof of Proposition 5. Consider the counter-example illustrated in Figure EC.3 for $\lambda_1 = 0.8, \lambda_2 = 0.2$. The MMDP has 5 states, 2 actions, 2 models, and 2 decision epochs. First, we can explicitly enumerate all possible deterministic policies for the non-adaptive weighted value problem.

Table EC.2 An explicit enumeration of the weighted value under every possible deterministic policy for the non-adaptive weighted value problem.

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

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

WSU will select $\pi(C)=1$ and update the value for each model as $v^1(C)=0$ and $v^2(C)=0$. For state B, WSU will select

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

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

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

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

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*

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

Proof of Lemma 1. For ease of notation, we refer to $\hat{\pi}(\lambda^1)$ as π^1. The value-to-go under policy π in model m from state s will be denoted as $v_t^m(s,\pi)$. Because $|\mathcal{M}|=2$, we will refer to the two models as m and $\bar{m}$ where λ_m is the weight on model m and $(1-\lambda_m)$ is the weight on model $\bar{m}$. Suppose the proposition is not true; that is, suppose there exists $\lambda_m^1>\lambda_m^2$ such that $v^m(\hat{\pi}(\lambda^1))<v^m(\hat{\pi}(\lambda^2))$. Then, it must be the case that for some $t\in\mathcal{T}$, $s\in\mathcal{S}$ that

$$v_t^m(s,\pi^1)<v_t^m(s,\pi^2). \tag{EC.4}$$

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

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

$$\begin{aligned}&\lambda_m^1 v_t^m(s,\pi^1)+(1-\lambda_m^1)v_t^{\bar{m}}(s,\pi^1)\geq\lambda_m^1 v_t^m(s,a)+(1-\lambda_m^1)v_t^{\bar{m}}(s,a), \quad \forall a\in\mathcal{A}\\ \Rightarrow&\lambda_m^1 v_t^m(s,\pi^1)+(1-\lambda_m^1)v_t^{\bar{m}}(s,\pi^1)\geq\lambda_m^1 v_t^m(s,\pi^2)+(1-\lambda_m^1)v_t^{\bar{m}}(s,\pi^2)\end{aligned} \tag{EC.5}$$

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

$$\begin{aligned}&\lambda_m^2 v_t^m(s,\pi^2)+(1-\lambda_m^2)v_t^{\bar{m}}(s,\pi^2)\geq\lambda_m^2 v_t^m(s,a)+(1-\lambda_m^2)v_t^{\bar{m}}(s,a), \quad \forall a\in\mathcal{A}\\ \Rightarrow&\lambda_m^2 v_t^m(s,\pi^2)+(1-\lambda_m^2)v_t^{\bar{m}}(s,\pi^2)\geq\lambda_m^2 v_t^m(s,\pi^1)+(1-\lambda_m^2)v_t^{\bar{m}}(s,\pi^1).\end{aligned} \tag{EC.6}$$

Rearranging (EC.5), we have

$$\lambda^1(v_t^m(s,\pi^1)-v_t^m(s,\pi^2))+(1-\lambda_m^1)(v_t^{\bar{m}}(s,\pi^1)-v_t^{\bar{m}}(s,\pi^2))\geq 0, \tag{EC.7}$$

and rearranging (EC.6), we have

$$\lambda_m^2(v_t^m(s,\pi^2)-v_t^m(s,\pi^1))+(1-\lambda_m^2)(v_t^{\bar{m}}(s,\pi^2)-v_t^{\bar{m}}(s,\pi^1))\geq 0 \tag{EC.8}$$

$$\Rightarrow -\lambda_m^2\left(v_t^m(s,\pi^1)-v_t^m(s,\pi^2)\right)-(1-\lambda_m^2)\left(v_t^{\bar{m}}(s,\pi^1)-v_t^{\bar{m}}(s,\pi^2)\right)\geq 0. \tag{EC.9}$$

Adding (EC.7) and (EC.9), we have:

$$\begin{aligned}&(\lambda_m^1-\lambda_m^2)\left(v_t^m(s,\pi^1)-v_t^m(s,\pi^2)\right)+\left((1-\lambda_m^1)-(1-\lambda_m^2)\right)\left(v_t^{\bar{m}}(s,\pi^1)-v_t^{\bar{m}}(s,\pi^2)\right)\geq 0\\ \Rightarrow&(\lambda_m^1-\lambda_m^2)\left(v_t^m(s,\pi^1)-v_t^m(s,\pi^2)+v_t^{\bar{m}}(s,\pi^2)-v_t^{\bar{m}}(s,\pi^1)\right)\geq 0.\end{aligned} \tag{EC.10}$$

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

$$\begin{aligned}&v_t^m(s,\pi^1)-v_t^m(s,\pi^2)+v_t^{\bar{m}}(s,\pi^2)-v_t^{\bar{m}}(s,\pi^1)\geq 0\\ \Rightarrow&v_t^{\bar{m}}(s,\pi^2)-v_t^{\bar{m}}(s,\pi^1)\geq v_t^m(s,\pi^2)-v_t^m(s,\pi^1)\\ \Rightarrow&v_t^{\bar{m}}(s,\pi^2)>v_t^{\bar{m}}(s,\pi^1),\end{aligned} \tag{EC.11}$$

where (EC.11) follows because of (EC.4). However, because $v_t^m(s,\pi^1)<v_t^m(s,\pi^2)$ and $v_t^{\bar{m}}(s,\pi^1)<v_t^{\bar{m}}(s,\pi^2)$, this implies that

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

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

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

□

Proposition 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.

Proof of Proposition 6. Let λ be the weight on model 1, π^1 be an optimal policy for model 1, and π^2 be an optimal policy for model 2. Due to the result of Proposition 1, it follows that

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

and therefore,

$$W(\hat{\pi}(\lambda)) = \lambda v^1(\hat{\pi}(\lambda)) + (1-\lambda)v^2(\hat{\pi}(\lambda)) \geq \lambda v^1(\pi^2) + (1-\lambda)v^2(\pi^1). \tag{EC.12}$$

Due to the upper bound discussed in Remark 4,

$$W(\hat{\pi}(\lambda)) \leq \lambda_1 v^1(\pi^1) + \lambda_2 v^2(\pi^2). \tag{EC.13}$$

From (EC.12) and (EC.13), the result follows. □

Proposition EC.1. *It is possible that there are no optimal policies that are Markovian for the adaptive problem.*

Proof of Proposition EC.1. Consider the MMDP illustrated in Figure EC.4.

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

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

Since state 4 is the only state in which there is a choice of action, we define the possible policies selecting an action in this state. Consider the adaptive problem for this MMDP. The optimal decision rule for state 4 will depend on the state observed at time $t = 2$: If the history of the MMDP is $(s_1 = 1, a_1 = 1, s_2 = 2, a_2 = 1)$, then select action 1, otherwise select action 2. In model 1, the only way to reach state 4 is through state 2. Upon observing this sample path, the policy prescribes taking action 1 which will lead to a transition to state 5 and thus a reward of 1 will be received. On the other hand, in model 2, the only way to reach state 4 is through state 3. Therefore, the

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

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

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

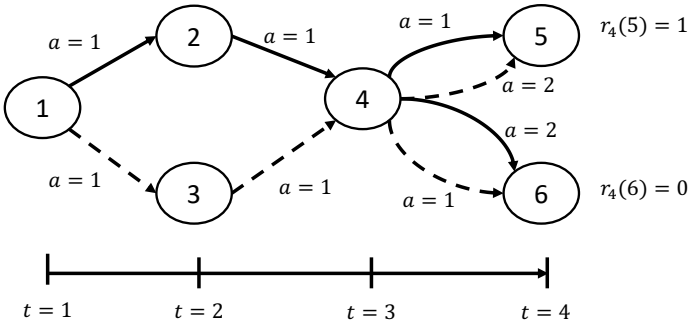


Figure EC.4 An example of an MMDP for $W_A > W_N$. The MMDP shown has six states, two actions, and two models. Each arrow represents a transition that occurs with probability 1 for the corresponding action labeling the arrow. Solid lines represent transitions in model 1 and dashed lines represent transitions in model 2. There are no intermediate rewards in this MMDP, but there is a terminal reward of 1 if state 5 is reached.

PROPOSITION EC.2. *Any MMDP can be recast as a special case of a POMDP such that the maximum weighted value of the MMDP is equivalent to the expected discounted rewards of the POMDP.*

Proof of Proposition EC.2. Let $(\mathcal{T}, \mathcal{S}, \mathcal{A}, \mathcal{M}, \Lambda)$ be an MMDP. From this MMDP, we can construct a POMDP in the following way. The *core states* of the POMDP will be constructed as

state-model pairs, $(s,m) \in \mathcal{S} \times \mathcal{M}$. The action space for the POMDP is the same as the action space for the MMDP, $\mathcal{A}$. We construct the rewards for the POMDP, denoted r^P, as follows:

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

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

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

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

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

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

Remark EC.2. If the state-model pairs that make up the POMDP core state space are ordered as $(1,1),\ldots,(S,1)$, $(1,2),\ldots,$ $(S,2),\ldots,(1,M),\ldots,(S,M)$, then the transition probability matrix has the following block diagonal structure:

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

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

Proposition EC.3. *The adaptive problem for MMDPs is PSPACE-hard.*

Proof of Proposition EC.3. This result follows from the original proof of complexity for POMDPs from Papadimitriou and Tsitsiklis (1987). Although the MMDP is a special case of a POMDP, we illustrate that the special structure in the observation matrix and transition probabilities is precisely the special case of POMDPs used in the original complexity proof. To aid

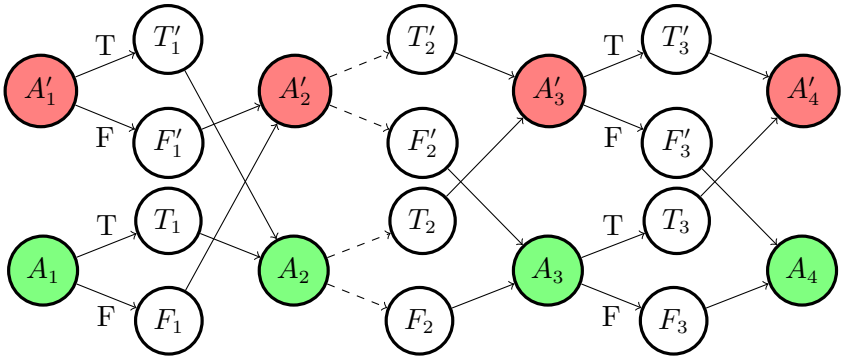
the reader's understanding, we reproduce the proof here with the modifications to make it specific to MMDPs. We also provide Figure EC.5 which illustrates the construction of an MMDP from the quantified satisfiability problem with two clauses for two existential variables and a universal variable.

First, we assume that $\lambda_m \in (0,1)\ \forall m \in \mathcal{M}$. To show that the adaptive weighted value problem for MMDPs is PSPACE-hard, we reduce QSAT to this problem. We start from any quantified boolean formula $(Q_1 u_1)(Q_2 u_2)\cdots(Q_n u_n)F(u_1, u_2, \ldots, u_n)$ with n variables, n quantifiers (i.e, Q_i is $\exists$ or $\forall$), and m clauses $C_1, C_2, \ldots, C_m$. We construct an MMDP with m models such that its optimal policy has weighted value of 0 or less if and only if the formula is true. The MMDP is constructed as follows: for every variable u_i, we will generate states corresponding to two decision epochs $2i-1$ and $2i$. In decision epoch $2i-1$, there will be two states, A'_i and A_i. In decision epoch $2i$, there will be four states, T'_i, F'_i, T_i, and F_i. After the last decision epoch (at time $2n+1$), there will be 2 states, A_{n+1} and A'_{n+1}. The initial state is A'_1 for every model. The action space is constructed as follows: for every existential variable u_i, the states A'_i and A_i each have two possible actions, *true* (T) and $false$ (F), which are elements of the action set $\{T, F\}$. All other states have only one action. The models of the MMDP correspond to the clauses in the quantified formula. Each model's transition probabilities are defined as follows: for every existential variable, the transitions out of A'_i and A_i are deterministic according to the action taken. For state A'_i (A_i), selecting action *true* will ensure that the next state is T'_i (T_i) and selecting action $false$ will ensure that the next state is F'_i (F_i). For every universal variable u_i, the transitions from A'_i (A_i) to T'_i (T_i) and from A'_i (A_i) to F'_i (F_i) occur with equal probability. The differences between the models' transition probabilities occur depending on the negation of variables within the corresponding clause. For every variable u_i that is not negated in the clause, transitions occur deterministically from T'_i to A_{i+1}, F'_i to A'_{i+1}, T_i to A_{i+1}, and F_i to A'_{i+1}. For every variable u_i that is negated in the clause, transitions occur deterministically from T'_i to A'_{i+1}, F'_i to A_{i+1}, T_i to A'_{i+1}, and F_i to A_{i+1}. The initial state is A'_1 for every model. There is a terminal cost of 1 upon reaching state A'_{n+1} and no cost for reaching A_{n+1}. Other than the terminal costs, there are no costs associated with any of the states or actions.

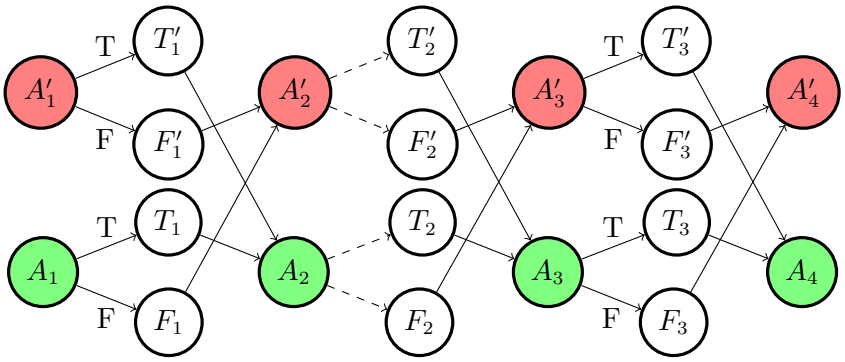
Now that we have constructed the MMDP, we must show that there exists a policy that achieves a weighted value of zero if and only if the statement is true. First, we show that if there exists a history-dependent policy with a weighted value of zero, then the statement must be true. Consider that such a policy exists. Recall that for every model, we start in state A'_1. In order to achieve a weighted value equal to zero, the policy must ensure that we end in state A_{n+1} for every model. If not, we incur a cost of 1 at time $2n+1$ in one of the models $m \in \mathcal{M}$ which has weight $\lambda_m > 0$, and thus the weighted value is not zero. If we were able to reach state A_{n+1} in every model, this would

imply that our policy is able to select actions for states A'_i and A_i for existential variables u_i based on observation of the previous universal variables in a way that the clause is satisfied. Since this occurs for all models, each clause must be true.

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



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



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

Figure EC.5 An illustration of how the quantified formula $\exists u_1 \forall u_2 \exists u_3 (u_1 \vee !u_2 \vee u_3) \wedge (u_1 \vee u_2 \vee !u_3)$ can be represented as an MMDP. Solid lines represent transitions that occur with probability. Dashed lines represent transitions that occur out of the state with equal probability. Transitions corresponding to the actions *true* and $false$ are labeled with T and F, respectively. State A'_i represents the case where the clause is false at this point and states A_i represents the case where the clause is true at this point.

PROPOSITION EC.4 *The information state, b_t, has the following properties:*

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

Proof of Proposition EC.4.1. We will prove this by induction. At time $T+1$, the value function is represented as

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

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

$$\begin{aligned}
v_t(b_t) &= \max_{a_t \in \mathcal{A}} \left\{ b'_t r_t(a_t) + \alpha \sum_{s_{t+1} \in \mathcal{S}} \gamma(s_{t+1}|b_t, a_t) v_{t+1}(T(b_t, a_t, s_{t+1})) \right\} \\
&= \max_{a_T \in \mathcal{A}} \left\{ b'_t r_t(a_t) + \alpha \left[\sum_{s_{t+1} \in \mathcal{S}} \left(\sum_{m' \in \mathcal{M}} \sum_{s_t \in \mathcal{S}} p^{m'}(s_{t+1}|s_t, a_t) b_t(s_t, m') \right) \cdot v_{t+1}(T(b_t, a_t, s_{t+1})) \right] \right\} \\
&= \max_{a_t \in \mathcal{A}} \left\{ \sum_{s_t \in \mathcal{S}} \sum_{m \in \mathcal{M}} r_t^m(s_t, a_t) \cdot b_t(s_t, m) + \alpha \sum_{s_{t+1} \in \mathcal{S}} \sum_{m \in \mathcal{M}} \max_{\beta_{t+1} \in B_{t+1}} \beta_{t+1}(s_{t+1}, m) \cdot \sum_{s_t \in \mathcal{S}} p^m(s_{t+1}|s_t, a_t) b_t(s_t, m) \right\} \\
&= \max_{a_t \in \mathcal{A}} \left\{ \sum_{s_t \in \mathcal{S}} \sum_{m \in \mathcal{M}} \left(r_t^m(s_t, a_t) + \sum_{s_{t+1} \in \mathcal{S}} \max_{\beta_{t+1} \in \mathcal{B}_{t+1}} \beta_{t+1}(s_{t+1}, m) \cdot p^m(s_{t+1}|s_t, a_t) \right) b_t(s_t, m) \right\},
\end{aligned} \tag{EC.14}$$

which is piece-wise linear and convex in b_t. Therefore, we can represent (EC.14) as the maximum over a set of hyperplanes:

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

where

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

□

Proof of Proposition EC.4.2 This follows directly from the definition of the information state EC.1 and the definition of the conditional probabilities in (C). To elaborate, we prove this by

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

Proof of Proposition EC.4.3 Next, we show that the information state can be efficiently transformed in each decision epoch using Bayesian updating. That is, we aim to show that the information state is Markov in that the information state at the next stage only depends on the information state in the current stage, the action taken, and the state observed in the next stage:

$$b_{t+1} = T(b_t, a_t, s_{t+1}) \tag{EC.15}$$

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

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

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

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

where

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

which follows from the following:

$$\begin{aligned}
b_{t+1}(s_{t+1}, m) &= \mathbb{P}(m \mid h_{t+1}) \\
&= \mathbb{P}(m \mid s_{t+1}, a_t, h_t) && \text{(EC.16)} \\
&= \frac{\mathbb{P}(m, s_{t+1} \mid a_t, h_t)}{\mathbb{P}(s_{t+1} \mid a_t, h_t)} && \text{(EC.17)} \\
&= \frac{\mathbb{P}(s_{t+1} \mid m, a_t, h_t)\, \mathbb{P}(m \mid a_t, h_t)}{\sum_{m' \in \mathcal{M}} \mathbb{P}(s_{t+1} \mid m, a_t, h_t)\, \mathbb{P}(m \mid a_t, h_t)} && \text{(EC.18)} \\
&= \frac{\mathbb{P}(s_{t+1} \mid m, a_t, h_t)\, \mathbb{P}(m \mid h_t)}{\sum_{m' \in \mathcal{M}} \mathbb{P}(s_{t+1} \mid m, a_t, h_t)\, \mathbb{P}(m \mid h_t)} && \text{(EC.19)} \\
&= \frac{\sum_{s_t \in \mathcal{S}} p_t^m(s_{t+1} \mid s_t, a_t) \mathbb{1}(s_t) \mathbb{P}(m \mid h_t)}{\sum_{m' \in \mathcal{M}} \sum_{s_t \in \mathcal{S}} p_t^{m'}(s_{t+1} \mid s_t, a_t) \mathbb{1}(s_t) \mathbb{P}(m' \mid h_t)} && \text{(EC.20)} \\
&= \frac{\sum_{s_t \in \mathcal{S}} p_t^m(s_{t+1} \mid s_t, a_t) b_t(s_t, m)}{\sum_{m' \in \mathcal{M}} \sum_{s_t \in \mathcal{S}} p_t^{m'}(s_{t+1} \mid s_t, a_t) b_t(s_t, m')}, && \text{(EC.21)}
\end{aligned}$$

if $s_{t+1} \in \mathcal{S}$ is in fact the state observed at time $t+1$. (EC.16) follows from the definition of h_{t+1}, and (EC.17) and (EC.18) follow from the laws of conditional probability and total probability. (EC.19) follows because the action is selected independently of the context. (EC.20) follows from the definition of $p^m(s_{t+1} \mid s_t, a_t)$ and an indicator which denotes the state at time t, and (EC.21) follows from the definition of the information state at time t. We define the operator T such that the element at (s, m) in $T(b_t, a_t, s_{t+1})$ is exactly $T^m(b_t, a_t, s_{t+1})$ if $s = s_{t+1}$ and 0 otherwise.

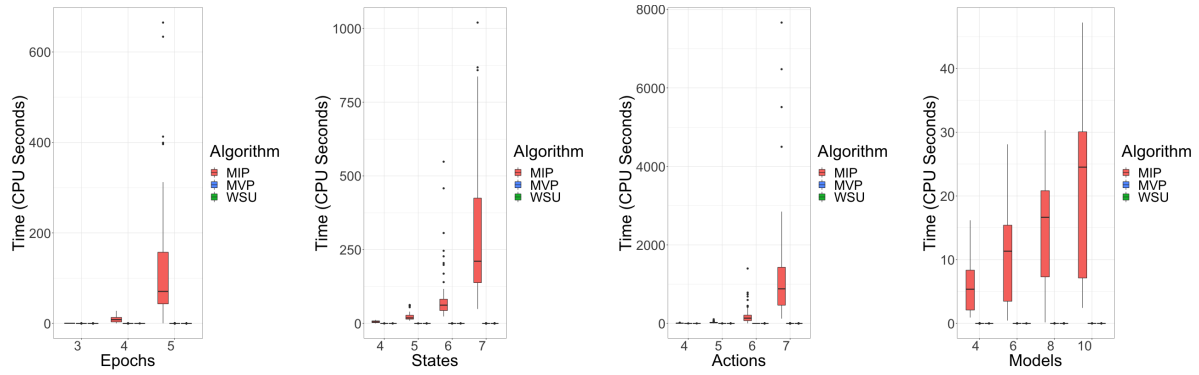
Therefore, the information state is Markovian in that the information state at time $t+1$ only relies on the information state at time t, the action taken at time t, and the state observed at time $t+1$. $\square$

Appendix D: Additional results and computational experiments

In this appendix, we provide more results from our computational experiments and introduce another set of test instances that we use to compare the Weight-Select-Update (WSU), maen value problem (MVP), and mixed-integer programming (MIP) solution methods. We also provide more results from the case study described in Section 7.

D.1. Additional results from the random instance computational experiments

Figure EC.6 demonstrates the run-time of the three proposed solution methods on the various sizes of the random test instances described in Section 6 in the main body.



(a) Decision epochs (b) States (c) Actions (d) Models

Figure EC.6 Boxplots showing the effect of the number of decision epochs, states, actions, and models on computation time in the random instances. A base case problem size of 4 states, 4 actions, 4 models, and 4 decision epochs was used. In (a), only the number of decision epochs was varied (from 3 to 5). In (b), only the number of states was varied (4 to 7). In (c), only the number of actions was varied (4 to 7). In (d), only the number of models was varied (from 4 to 10).

Each algorithm was run on a set of 90 instances of the corresponding problem size. We observe that computation time to solve the MIP increases most quickly with respect to the number of decision epochs than the number of models. However, we see that the computation times required for the WSU and MVP heuristics increase at a much slower rate.

D.2. Additional computational experiments

We now describe a second set of computational experiments comparing the WSU, MVP, and MIP solution methods. These test instances are based on a small MDP for determining the most cost-effective HIV treatment policy (Chen et al. 2017).

D.2.1. Test instances We also consider a second set of test instances which matches the medical decision making context of our case study. The example we consider has been used many times in the medical decision making literature for illustrative purposes to demonstrate various methods.

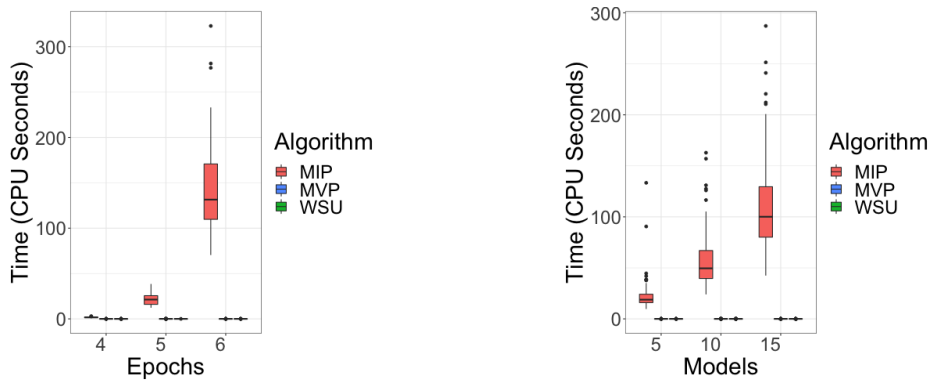
In this set of experiments, we consider an MDP for determining the optimal timing of treatments for HIV. In the MDP, HIV is characterized according to 4 health states: Mild, Moderate, Severe, or Dead. The patient transitions from the less severe states to the more severe states according to a Markov chain. The DM can choose to start the patient on one of three treatments: Treatment A, Treatment B, and Treatment C. Treatment A is the least effective but also the least expensive while Treatment C is the most effective but comes at the highest cost. Chen et al. (2017) provides a summary table of parameter values for this MDP as well as some sampling distributions for each parameter. In our experiments, we construct an MMDP by sampling parameters from the corresponding distributions. We consider 10, 20, and 30 models in the MMDP and vary the number of decision epochs from 5 to 10 to explore how the proposed methods perform.

D.2.2. Results Figure EC.7 demonstrates the run-time of the three proposed solution methods on the medical decision making instances. We find that the MVP and WSU were able to solve these instances relatively quickly (under 0.1 CPU seconds for each instance) while the average time to solve the MIP noticeably increases as the size of the number of decision epochs increases (from 1.73 CPU seconds on average for 4 decision epochs to 141.84 CPU seconds for 6 decision epochs). For the instances with 6 decision epochs, the MIP computation time rose from 21.73 CPU seconds on average for 5 models to 111.04 CPU seconds for 15 models. Comparing WSU and MVP in terms of optimality gap, we observe that for these test instances, both WSU and MVP perform quite well with maximum optimality gaps under 0.45% and 0.69% respectively. These results suggest that the MVP and WSU heuristics may be suitable for generating solutions to medical decision making instances. The case study in Section 7 considers a larger medical decision making problem in the context of preventive blood pressure and cholesterol management.

D.3. Additional results from the case study of blood pressure and cholesterol management

We now discuss the policies associated with the solution generated using WSU when the weights are treated as an uninformed prior on the models for the case study described in Section 7.

Figures EC.8(a) and EC.8(b) illustrate medication use for male and female patients, respectively, under three different policies: the ACC/AHA model's optimal policy, the FHS model's optimal policy, and a policy generated via WSU with $\lambda_F = \lambda_A = 50\%$. These figures illustrate the probability that a patient who follows the specified policy from age 54 will be on the corresponding medication, conditioned on the patient being alive, as a function of their age. For men, the optimal policy for FHS model and the optimal policy for the ACC/AHA model agree that all men should start statins

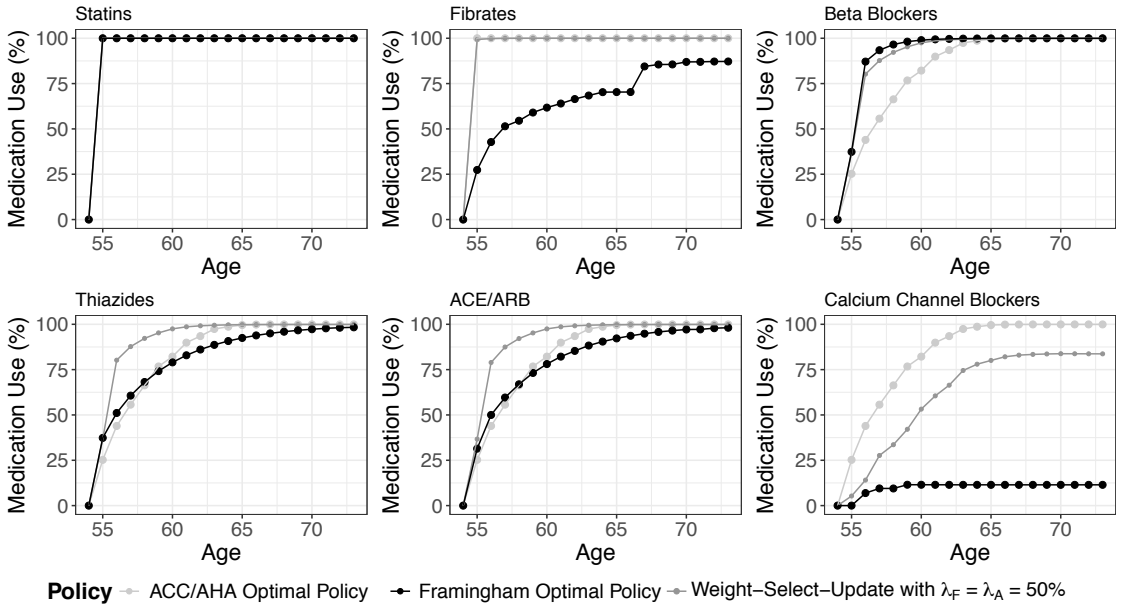


(a) Computation time vs. number of decision epochs

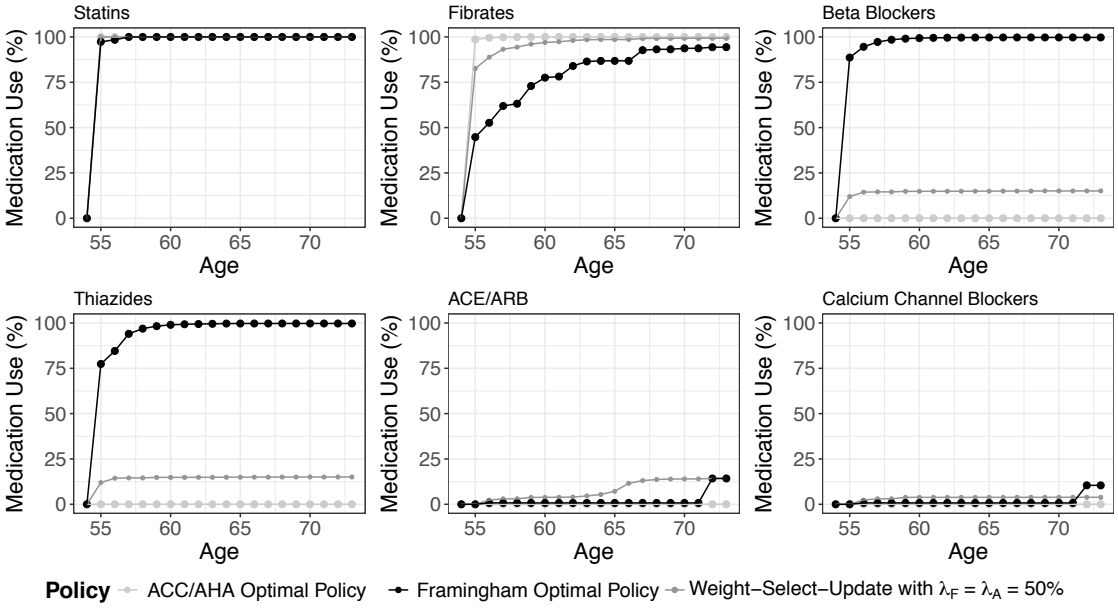
(b) Computation time vs. number of models

Figure EC.7 Boxplots showing the effect of the number of decision epochs and number of models on computation time in the medical decision making instances. Each algorithm was run on a set of 100 instances of the corresponding problem size. We observe that computation time to solve the MIP increases more quickly with respect to the number of decision epochs than the number of models. However, we see that the computation time required for the WSU and MVP heuristics increases at a much slower rate.

immediately, which could be explained by the relatively low disutility and high risk reduction of statins in both models. However, the models disagree in the use of fibrates and the 4 classes of blood pressure medications. The optimal policy for the ACC/AHA model suggests that all men should start fibrates immediately, suggesting that cholesterol control is important in the ACC/AHA model. However, fibrates are less commonly prescribed under the FHS model's optimal policy with about two-thirds of men on this medication by age 65. The policy generated with WSU agrees with the ACC/AHA policy's more extensive use of fibrates which may suggest that focusing on cholesterol control could be a good strategy in both models. Among the blood pressure medications, there are some disagreements between the optimal policies of the two models, with the most distinct being for the use of calcium channel blockers. This is likely to be due to the relatively high disutility (from side effects of calcium channel blockers) and low risk reduction associated with this medication. In the ACC/AHA model, the risk reduction of calcium channel blockers is worth the disutility in many cases, but in the FHS model, there are few instances in which the disutility associated with this medication is worth the gain in QALYs. The policy generated with WSU generates a policy that strikes a balance between these two extremes. While the differences are not quite as extreme, WSU also generates a policy that balances the utilization of thiazides prescribed by each model's optimal policy. For the other classes of blood pressure medications, both models agree that these medications should be commonly used for men, but disagree in the prioritization of these medications. The ACC/AHA model tends to utilize these medications more at latter ages,



(a) Male



(b) Female

Figure EC.8 The percentage of patients who have not died or had an event by the specified age that will be on a medication under each of three different treatment policies: the ACC/AHA model's optimal policy, the FHS model's optimal policy, and a policy generated via WSU with $\lambda_F = 50\%$, as evaluated in the FHS model.

while the FHS model starts more men on these medications early. Interestingly, WSU suggests that starting ACE/ARBs and beta blockers earlier is a good strategy in both models.

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

Appendix E: Case study with multiple natural history models

In this appendix, we provide an example of the case study presented in Section 7 but modified to include more than two models. As in Section 7, the MMDP presented here includes two possible models for cardiovascular risk. However, we now also consider three natural history models of blood pressure and cholesterol progression. The two models of cardiovascular risk are the FHS risk model (Wolf et al. 1991, Wilson et al. 1998) and the ACC/AHA risk model (Goff et al. 2014), as before. These are the most well-known risk models used by physicians in practice. The three models of blood pressure and cholesterol progression include the maximum likelihood estimate from longitudinal data plus two other scenarios with quicker and slower progression. We may view the scenarios to be based on sensitivity analysis relative to our base case model adapted from Mason et al. (2014). The "Basecase Progression" model was estimated from empirical data of Denton et al. (2009). Figure EC.9 illustrates how the "Quick Progression" and "Slow Progression" scenarios relate to the basecase scenario. In the "Quick Progression" scenario, all transition probabilities that corresponding to a worsening of the health condition (e.g., low SBP to high SBP) are scaled by $\beta = 150\%$ and all transitions corresponding to an improvement of health (e.g., high SBP to low SBP) are scaled by $\alpha = 50\%$. Analogously, in the "Slow Progression" scenario all transition probabilities corresponding to a worsening of the health condition are scaled by $\beta = 50\%$ and improving transitions are scaled by $\alpha = 150\%$. These two scenarios – slow and quick progression – may be viewed as models defined by different patient groups with differing physiological factors or phenotypes governing disease progression. We create each model of the final MMDP by selecting one of the two cardiovascular risk models and one of the nature history models. Therefore, this version of the MMDP has 4099 states, 64 actions, 20 decision epochs, and 6 models each with equal weight.

E.1. Results

Using the MMDP described above, we evaluated the performance of the WSU policy relative to each individual models' policy as well as a Rectangular Max-Min (RMM) formulation. The RMM is the classical max-min finite scenario model of Nilim and El Ghaoui (2005) wherein the MMDP is projected onto an (s, a)-rectangular ambiguity set. More information about the RMM formulation can be found in Appendix E.2.

Figure EC.10 shows the weighted QALYs gained relative to no treatment for the WSU heuristic and for the individual policies corresponding to each model of the MMDP. These results show that using the WSU heuristic to generate a solution to the WVP can provide a policy that performs better than each of the policies found by independently solving individual models of the MMDP.

Figure EC.11 shows the performance of the WSU policy and the RMM policy in terms of the weighted value across the models and also in the worst-case model for each policy. Although

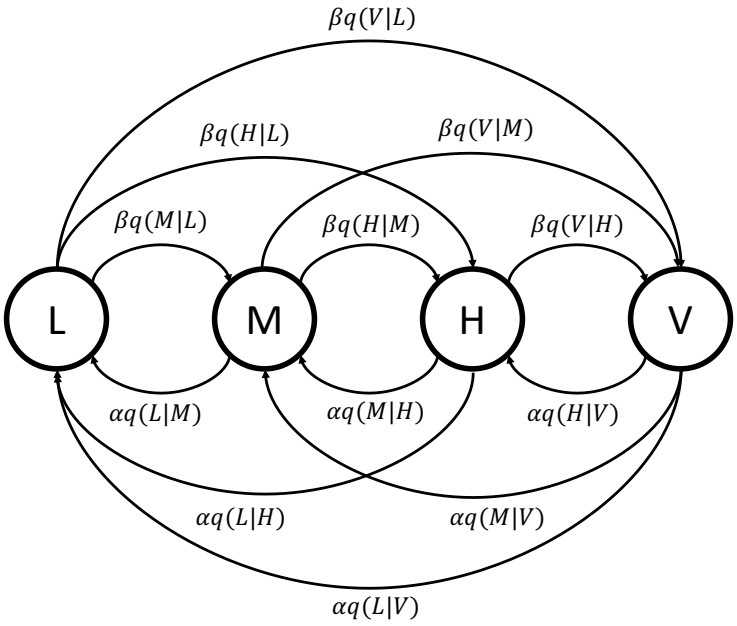
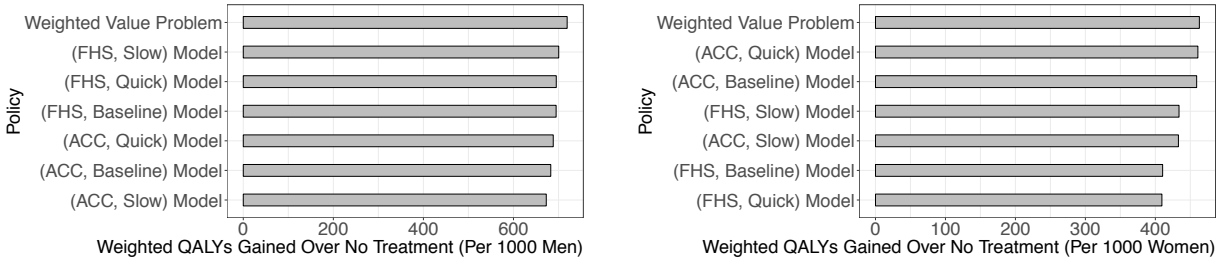


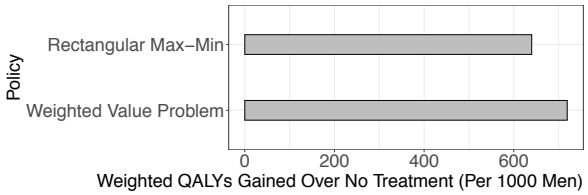
Figure EC.9 An illustration of the natural history models used in the cardiovascular disease (CVD) case study.



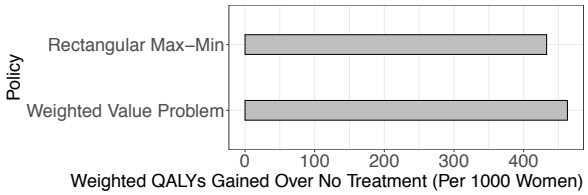
(a) Male (b) Female

Figure EC.10 The weighted QALYs gained relative to no treatment (reported as QALYs per 1000 persons) reported for each of the individual models' optimal policies and the policy from the WSU heuristic.

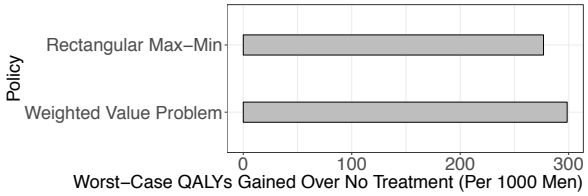
the WSU is not explicitly accounting for worst-case outcomes, we observe that the WSU policy outperforms the RMM policy in terms of worst-case. The reason for this is that, by projecting the MMDP onto a rectangular ambiguity set in RMM, the DM has to protect against a worst case in which the model of the MDP is allowed to change across states, actions, and decision-epochs. The worst case that the DM is protecting against is not a realistic representation of what is happening in the worst-case model and so this policy does not perform well. The WSU policy also performs better in terms of the weighted value case. These findings suggests that a weighted approach to parameter ambiguity might outperform a rectangular ambiguity set approach.



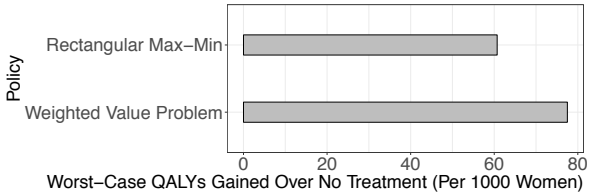
(a) Male, Weighted QALYs



(b) Female, Weighted QALYs



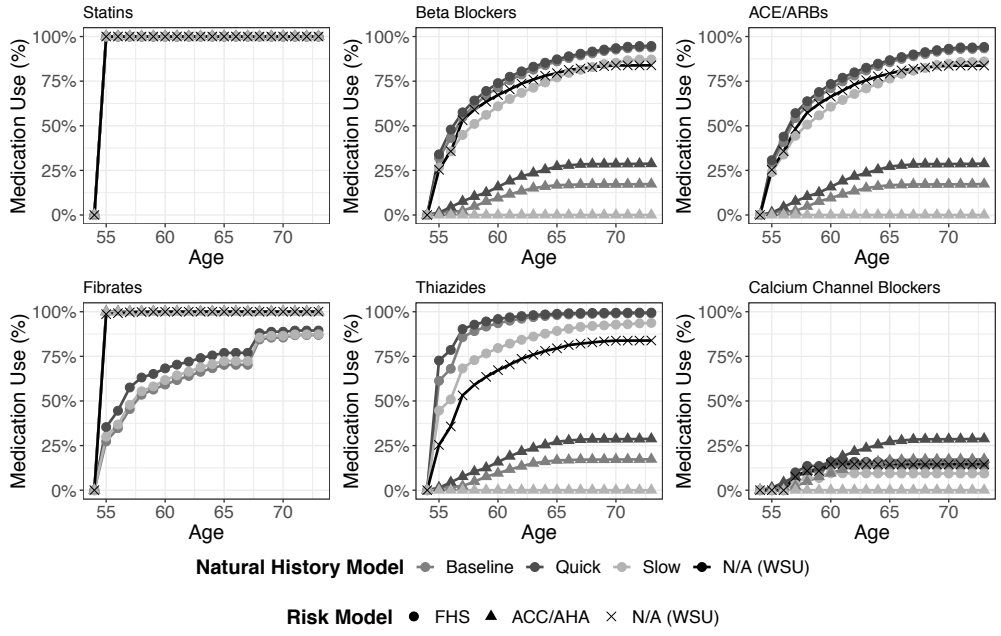
(c) Male, Worst-case QALYs



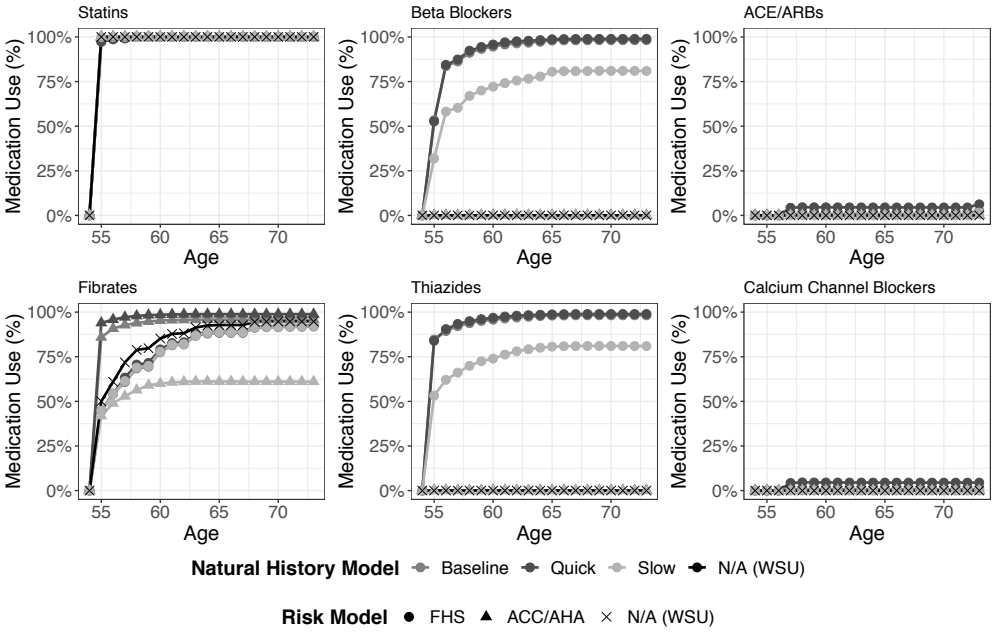
(d) Female, Worst-case QALYs

Figure EC.11 The weighted QALYs gained relative to no treatment over all 6 models and worst-case QALYs gained relative to no treatment over all 6 models (reported as QALYs per 1000 persons) for the weighted value problem policy (found with WSU) and the RMM policy.

Figure EC.12(a) illustrates the percentage of men that would be on a given medication by a particular age under the WSU heuristic's policy and under each policy found by solving an individual model of the MMDP. The lines with circular markers represent the optimal policies corresponding to each model that considers the FHS risk model and the lines with triangular markers represent the optimal policies corresponding to each model that considers the ACC/AHA risk model. The color indicates which model of the transitions among the health states was used. The black line represents the medication usage corresponding to the WSU policy. For cholesterol control, there is little disagreement about whether or not men should start statins right away which is consistent with findings in the medical literature that suggest statins are highly effective at lowering risk of cardiovascular events. There is more disagreement on the use of fibrates. The ACC/AHA models all suggest that all men should start fibrates immediately, but the FHS models indicate that not all men should start this drug by age 75. The WSU policy agrees with the ACC/AHA models in this case. In terms of blood pressure medication use, the WSU policy recommends initiation of beta blockers and ACE/ARBS that is similar to that suggested by the FHS models. However, the WSU policy suggests lower levels of thiazide initiation that the FHS models. Each of the policies agree that calcium channel blockers should be not be used as often which makes sense given that these drugs are the least effective in this model with similar disutility. Relative to each of the models, the WSU policy suggests a similar strategy for cholesterol control as the ACC/AHA models while more aggressive use of thiazides than the FHS models. For most blood pressure medications, the initiation suggested by WSU is in line with that suggested by the FHS models and more aggressive than that suggested by the ACC/AHA models.



(a) Male



(b) Female

Figure EC.12 The percentage of patients who have not died or had an event by the specified age that will be on a medication under different policies. We consider each model's individual policy and the policy generated via WSU. The color of the line indicates the natural history model was used and the marker indicates the risk model. WSU is denoted by the black line with the X marker. The medication usage is evaluated in the (FHS, Baseline) model.

Figure EC.12(b) illustrates medication usage for women. Overall, the WSU policy is more aggressive than the optimal policies for the FHS models and the slow progression ACC/AHA models in terms of cholesterol control. However, the WSU agrees with the ACC/AHA policies in terms of low initiating rates of blood pressure medications. The reason for this is that, for many women, blood pressure medications lead to side effects without reducing risk of CVD events in the ACC/AHA model. Therefore, using no medication appears like a better treatment option in this model. WSU protects against this by not selecting an action that would decrease the weighted value for a particular state.

In summary, the results of this case study illustrate how the policy generated by WSU trades off performance with respect to multiple models of CVD including two statistical models of risk of heart events (ACC/AHA and FHS) and three different models of the natural history of blood pressure and cholesterol. Our findings suggest that the WSU heuristic recommends less aggressive cholesterol treatment for men and more aggressive use of the blood pressure medication class of thiazides than the ACC/AHA guidelines. For women, there are more serious conflicts concerning the recommendations for the models indicating that blood pressure therapy appears beneficial in the FHS models but harmful when using the ACC/AHA risk estimator to estimate risk. Therefore, the WSU policy recommends the use of cholesterol medications rather than blood pressure medications for women. This information could be useful for informing policy-makers who are tasked with designing screening and treatment protocols in the face of conflicting information from the medical literature.

E.2. Rectangular Max-Min Formulation

In this section, we provide more detail about the RMM formulation. Procedure 4 solves a robust MDP formulation of the MMDP using the finite scenario model describe in Nilim and El Ghaoui (2005). To guarantee a tractable robust MDP formulation, we employ the commonly used (s,a)-rectangularity property which imposes independence between rows of the transition probability matrix. To satisfy the (s,a)-rectangularity property, we project the parameters in the MMDP onto an (s,a)-rectangular ambiguity set. The projection is done by constructing a ambiguity set that is independently constructed for each (s,t,a)-tuple for $(s,t,a) \in \mathcal{S} \times \mathcal{T} \times \mathcal{A}$:

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

and

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

with

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

Procedure 4 The RMM algorithm.

Input: MMDP

Let $v_{T+1}^{WC}(s_{T+1}) = \min_{m\in\mathcal{M}}\{r_{T+1}^m(s_{T+1})\}$

$t \leftarrow T$

while $t \geq 1$ **do**

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

$$\pi_t^{WC}(s_t) \leftarrow \arg\max_{a_t\in\mathcal{A}} \left\{ \min_{m\in\mathcal{M}} \left(r_t^m(s_t,a_t) + \sum_{s_{t+1}\in\mathcal{S}} p_t^m(s_{t+1}|s_t,a_t) v_{t+1}^{WC}(s_{t+1}) \right) \right\}$$

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

end for

$t \leftarrow t-1$

end while

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

and

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

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

$$\max_{\pi\in\Pi} \min_{P\in\mathscr{P}, R\in\mathscr{R}} \mathbb{E}^{\pi,P,R} \left[\sum_{t=1}^{T} r_t(s,\pi_t(s)) + r_{T+1}(s) \right]. \tag{EC.22}$$

By construction, the ambiguity set has the (s,a)-rectangularity property so (EC.22) can be solved efficiently using Procedure 4.

References

Boloori A, Saghafian S, Chakkera HA, Cook CB (2020) Data-driven management of post-transplant medications: an ambiguous partially observable Markov decision process approach. *Manufacturing & Service Operations Management* 22(5):1066–1087.

Brunskill E, Li L (2013) Sample complexity of multi-task reinforcement learning. *International Conference on Uncertainty in Artificial Intelligence (UAI)* 122–131.

Chen Q, Ayer T, Chhatwal J (2017) Sensitivity analysis in sequential decision models: a probabilistic approach. *Medical Decision Making* 37(2):243–252.

Delage E, Mannor S (2009) Percentile Optimization for Markov Decision Processes with Parameter Uncertainty. *Operations Research* 58(1):203–213.

Denton BT, Kurt M, Shah ND, Bryant SC, Smith SA (2009) Optimizing the Start Time of Statin Therapy for Patients with Diabetes. *Medical Decision Making* 29(3):351–367.

Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, Wilson PWF (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.

Goyal V, Grand-Clement J (2018) Robust Markov decision process: Beyond rectangularity. *arXiv preprint arXiv:1811.00215* .

Hallak A, Di Castro D, Mannor S (2015) Contextual Markov decision processes, URL `http://arxiv.org/abs/1502.02259`.

Iyengar GN (2005) Robust dynamic programming. *Mathematics of Operations Research* 30(2):257–280.

Kaufman DL, Schaefer AJ, Roberts MS (2011) Living-donor liver transplantation timing under ambiguous health state transition probabilities – Extended abstract. *Proceedings of the 2011 Manufacturing and Service Operations Management (MSOM) Conference*.

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

Li X, Zhong H, Brandeau ML (2017) Quantile markov decision process URL `http://arxiv.org/abs/1711.05788`.

Mannor S, Mebel O, Xu H (2016) Robust MDPs with k-Rectangular Uncertainty. *Mathematics of Operations Research* 41(4):1484–1509.

Mason JE, Denton BT, Shah ND, Smith SA (2014) Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients. *European Journal of Operational Research* 233(3):727–738.

Nilim A, El Ghaoui L (2005) Robust control of Markov decision processes with uncertain transition matrices. *Operations Research* 53(5):780–798.

Papadimitriou C, Tsitsiklis J (1987) The complexity of Markov decision processes. *Mathematics of Operations Research* 12:441–450.

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

Scheftelowitsch D, Buchholz P, Hashemi V, Hermanns H (2017) Multi-objective approaches to Markov decision processes with uncertain transition parameters. *Proceedings of the 11th EAI International Conference on Performance Evaluation Methodologies and Tools*, 44–51.

Sinha S, Kotas J, Ghate A (2016) Robust response-guided dosing. *Operations Research Letters* 44(3):394–399.

Smallwood RD, Sondik EJ (1973) The optimal control of partially observable Markov decision processes over a finite horizon. *Operations Research* 21(5):1071 – 1088.

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

Wiesemann W, Kuhn D, Rustem B (2013) Robust Markov decision processes. *Mathematics of Operations Research* 38(1):153–183.

Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. *Circulation* 97(18):1837–1847.

Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB (1991) Probability of stroke: A risk profile from the Framingham Study. *Stroke* 22(3):312–318.

Xu H, Mannor S (2012) Distributionally robust Markov decision processes. *Mathematics of Operations Research* 37(2):288–300.

Zhang Y, Steimle LN, Denton BT (2017) Robust Markov decision processes for medical treatment decisions 37, URL `http://www.optimization-online.org/DB_FILE/2015/10/5134.pdf`.