Multi-model Partially Observable Markov Decision Processes

Weiyu Li[a], Brian T. Denton[a,*]

[a]*Department of Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109, USA*

Abstract

We propose a new multi-model partially observable Markov decision process (MPOMDP) model to address the issue of model ambiguity in partially observable Markov decision process. Here, model ambiguity is defined as the case where there are multiple credible optimization models with the same structure but different model parameters. The proposed MPOMDP model aims to learn the distribution of the true model from system outputs over time, and to find the single optimal policy that maximizes the expected sum of all future rewards in all possible models. We discuss important structural properties of the proposed MPOMDP model, which not only reveal the benefit of the MPOMDP model by accounting for model ambiguity, but also motivate solution methods for MPOMDP. We develop an exact solution method, and two approximation methods that are shown to converge asymptotically, and compare their performance in computational experiments. Lastly, we use a case study of prostate cancer active surveillance to demonstrate how the MPOMDP model can be applied to a real-world problem to improve medical decision-making by created policies that are robust to different parameters in the multiple plausible models.

Keywords: multi-model partially observable Markov decision processes, stochastic optimization, model ambiguity, prostate cancer active surveillance

1. Introduction

Partially observable Markov decision process (POMDP) has been found successful in many problems, including machine maintenance, robot navigation, healthcare, and others (see Cassandra (1998) for a survey). This paper addresses the issue of *model ambiguity* in POMDP models defined as follows. In a POMDP model, the decision-maker can take actions to influence the transition dynamic, output, and reward from the system, such that the expectation of all future rewards is

*Corresponding author

Email addresses: weiyuli@umich.edu (Weiyu Li), btdenton@umich.edu (Brian T. Denton)

maximized. The transition, observation, and reward dynamics of a POMDP model are described by its model parameters. In practice, these model parameters are often estimated by pre-studies that fit models on historical observational data. A potential issue of this approach is that different studies can give different estimates of the model parameters. The difference in parameter estimates can arise from differences in the underlying study samples, study designs, model formulations, or other factors. In this study, we call it the issue of *model ambiguity* in POMDP models.

In this paper, we propose a new multi-model partially observable Markov decision process (MPOMDP) model to tackle the issue of model ambiguity. An MPOMDP model is a stochastic optimization and dynamic programming model that simultaneously considers multiple POMDP models, which have the same model structure but different model parameters. The goal is to find a single optimal policy that adaptively optimizes a "weighted" average of the value functions of all POMDPs. The model weight is given by the model belief vector, which can be interpreted as the importance and/or the probability of being the true model for each POMDP model, and is learned every time using the information from system outputs. In particular, even if none of the POMDPs considered in the MPOMDP is the true model for the study object, the learned belief is still guaranteed to assign a higher weight to the model with a greater probability of generating the observed outputs. Traditionally, when it comes to the issue of model ambiguity, a decision-maker may randomly pick a single model, take the average of multiple sets of model parameters, or conservatively consider a max-min model address the worst-case performance. In this study, we will show that the proposed MPOMDP model outperforms the traditional methods by achieving a non-negligible value of the stochastic solution (VSS) as defined in Birge (1982). Our study also sheds light on the expected value of perfect information (EVPI) (Schlaifer & Raiffa, 1961), which may be relevant in situations where there are opportunities to collect additional information to resolve model uncertainty.

We describe several important properties of the proposed MPOMDP model, which give insights into the model and motivate fast approximation methods we propose to solve the MPOMDP model. First, we show that an MPOMDP model can be reformulated as a special-case of the POMDP model, with an enlarged state space, thus inheriting many properties of the standard POMDP. Next, we discuss the existence and structure of the optimal policy of an MPOMDP model. Then, we describe solution methods that exploit properties of the model. Finally, we provide examples to illustrate the practical benefits of the proposed MPOMDP model.

The work of this paper is motivated by a healthcare application in prostate cancer active surveil-

lance (AS), which will be discussed in detail in Section 6. A pre-study by Li et al. (2023) developed a POMDP model to find the optimal timing for biopsies in prostate cancer AS, such that the burden of biopsy and the delay in detecting cancer progression are minimized. Their work in Li et al. (2020) first estimated the cancer progression rates, biopsy under-sampling errors, and PSA distributions using an HMM in four major prostate cancer AS studies in the world, which include the JH hospital, the UCSF medical center, the U of T medical center, and the PRIAS project. The study showed the model parameters to be statistically significantly different across studies. Motivated by this discrepancy, we consider the case of a new patient or new study for which the true model is unknown. We use computational experiments in Section 6 to show that our proposed MPOMDP model can find a single policy with the same complexity as the one given by a POMDP model, that achieves better overall performance based on clinical outcomes.

Our proposed MPOMDP model in this paper will mainly focus on the finite-horizon problem for several reasons. First, finite-horizon models are preferred over infinite-horizon models in healthcare applications and other applications where the survival time (length of decision epochs) can not be infinite and the model parameter can be non-stationary. Second, although a finite-horizon POMDP model can be easily reformulated as an infinite-horizon POMDP model by appending the time index to the state definition, it does not automatically solve the problem as the computational complexity would increase along with the size of the state space.

The rest of this article is organized as follows. In Section 2, we review the most related work in stochastic sequential decision-making under uncertainty and with model ambiguity, and summarize the main contribution of this work. In Section 3, we formally define the MPOMDP model. Next, in Section 4, we discuss some important structural properties of the MPOMDP. In Section 5, we present solution methods tailored to the MPOMDP model. We present the results of a toy example to demonstrate the computational properties of the proposed methods and a detailed case study of prostate cancer AS in Section 6. Finally, we conclude with a discussion of potential future research in Section 7.

2. Literature Review

In this section, we first review the most closely related work in sequential decision-making under uncertainty and model ambiguity. Then, we describe the main contributions of this paper with respect to the related literature.

The POMDP was first introduced by Åström (1965); Drake (1962) and Smallwood & Sondik

(1973). The POMDP model is a dynamic programming model for sequential decision-making, where the underlying system can be described by a hidden Markov model (HMM) (Rabiner & Juang, 1986). The objective of a POMDP model is to find the policy for actions to take at all time periods, such that the optimal cumulative reward is achieved. On the one hand, the POMDP model subsumes the HMM in that it adds decision-making about what action to take at each time period, which will influence the transition, output, and reward dynamics of the system. On the other hand, the POMDP model is a generalization of the Markov decision process (MDP) model (Puterman, 2014), where the underlying state is not observable and can only be inferred by the output of the system. POMDP models have found success in many problems, including machine maintenance (Ross, 1971), robot navigation (Cassandra et al., 1996), healthcare (Ayer et al., 2012; Zhang et al., 2012; Erenay et al., 2014), and many others (see Cassandra (1998) for a survey).

When applying the POMDP model to real-world problems, it is necessary to estimate model parameters that include the initial distribution function, transition probabilities, observation probabilities, and reward function. However, the estimation error and heterogeneity between different studies can induce ambiguity in the underlying HMM model. Li et al. (2023) used POMDP models to optimize AS strategies in prostate cancer, and showed that the optimal policies could vary considerably in different medical studies because of the difference in system dynamics revealed by model parameters.

Saghafian (2018) proposed an ambiguous POMDP (APOMDP) model to address the issue of model ambiguity in the POMDP model. Boloori et al. (2020) then applied the APOMDP model in a study of post-transplant medication management, which improved the existing policies by considering variability among physicians' attitudes toward ambiguous outcomes and patients' progression dynamics. In contrast to the work in this article, in their proposed APOMDP model, the objective function is in a robust optimization setting, which weights the best-case and worst-case value functions across different sets of model parameters. Moreover, they assumed that the best and worst models were selected independently over time, which might violate the Markov property and induce inconsistency in model dynamics across decision epochs. Nakao et al. (2021) described a distributionally robust Partially Observable Markov Decision Process (DR-POMDP), which estimates the distribution of the transition-observation probabilities using side information at the end of each time period, to maximize the worst-case reward for any joint-distribution of the ambiguous model parameters. In contrast, the study in this article seeks a single optimal policy that works well "on average", rather than optimizing the worst-case performance, when there are multiple credible

POMDP models.

Despite the short history of the study of model ambiguity in POMDP models, there has been a stream of research on model ambiguity in MDP models over the last two decades. Nilim & El Ghaoui (2005) and Iyengar (2005) considered a robust formulation of an MDP to optimize the worst-case performance (referred to as the "max-min" problem) of the model, while assuming a "rectangularity" property in ambiguity sets, i.e., the ambiguity in transition probabilities is independent with action, state, or time. They discussed the policy evaluation and other improved solution methods to the proposed robust MDP. Followed by their study, much of the research has focused on ways to construct ambiguity sets, to mitigate the rectangularity assumption on the ambiguity set, and to generalize the "max-min" objective function (Delage & Mannor, 2010; Xu & Mannor, 2012; Wiesemann et al., 2013; Delage & Iancu, 2015; Mannor et al., 2016). In contrast to these studies, our work in this paper addresses the issue of model ambiguity in a different manner. The MPOMDP model we proposed considers a weighted sum of value functions under different sets of model parameters, where the objective is to find a single policy that performs well overall possible models. Compared with the robust optimization formulation, our MPOMDP finds a less conservative policy that achieves the maximum of a weighted (by model belief) value function instead of the maximum worst-case value function.

The closest research to ours that we are aware of is that of Steimle et al. (2021), which formulated a multi-model Markov decision process (MMDP). They considered discrete ambiguity sets for the model parameters in MDPs with the objective of optimizing the weighted value function. They showed that any MMDP could be recast as a special case of a POMDP. Different from their study, our work in this paper considers a more complex setting of MPOMDP, where each single model is already a POMDP, so that we are encountered with much larger state and policy spaces (i.e., more severe curse of dimensionality and curse of history).

To close this section, we describe the main contributions of this paper to the literature. Our article is the first work addressing the issue of model ambiguity under the POMDP framework using the MPOMDP. In contrast to the work by Saghafian (2018), Nakao et al. (2021), and related literature on robust MDPs, our model formulation considers the objective function to be a weighted sum of value functions given by the belief vector under different sets of model parameters. Our formulation allows inter-dependent model transition, observation, and reward dynamics over time. Moreover, it provides less conservative policies than the robust optimization formulation, which aims to optimize the worst-case performance. Second, we study the structural properties of the

proposed MPOMDP, which not only motivate the solution methods, but also help analyze the effect of model ambiguity in POMDPs. Third, we describe an exact solution method, and two different approximation methods to our model that are shown to converge asymptotically and can provide near-optimal solutions in real-time. Finally, we present a case study for prostate cancer AS optimization, which illustrates how the MPOMDP can be applied in a real-world problem, and the benefit of the MPOMDP in stochastic sequential decision-making under model ambiguity.

3. Model Formulation

We start with a review of the formal definition of the POMDP, and then introduce the MPOMDP, which generalizes the POMDP for model ambiguity.

3.1. POMDP Definition

The POMDP model can be defined as follows:

Definition. A finite-horizon POMDP model $\mathcal{M}$ can be defined by a tuple

$$(S, b_0, A, P, O, F, r, T),$$

where S is the set of all states, b_0 is the initial distribution function over the set of states S, A is the set of all actions, $P : S \times A \times S \to [0, 1]$ is the state transition probability distribution, O is the set of all observations, $F : S \times A \times O \to [0, 1]$ is the observation probability distribution, $r : S \times A \times O \to \mathbb{R}$ is the reward function, and T is the length of time horizon.

Notice that in Definition 3.1, the state transition probability distribution, observation probability distribution, and the reward function are all stationary, i.e., independent of time. The definition of a non-stationary model can be easily adapted using time-dependent model parameters in finite-horizon POMDP formulation.

POMDP models are widely used to solve stochastic sequential decision-making problems with partially observable states. In a finite-horizon POMDP model, we can use $t = 0, 1, ..., T$ to denote its discrete time periods (also referred to as decision epochs), and b_t to denote the probability distribution over S (also referred to as a belief vector) at time $t \leq T$. Then, given a policy $\pi = (\pi_0, ..., \pi_T)$, where each π_t is a mapping from the space of the belief vector to A specifying the action to choose for all possible belief states at time t, the value function of the policy π starting from belief state b at time t is defined as

$$V_t^\pi(b_t) := \mathbb{E}^\pi[\sum_{k=t}^{T} \gamma^{t-k} r(s_k, a_k, o_k)|b_t], \ \forall b_t, \forall t \leq T,$$

where $\gamma \in [0,1]$ is a discount factor that diminishes the future rewards, s_k, a_k, and o_k are the state, action, and observation at time $k \leq T$, and the expectation is taken over all possible state, action, and observation trajectories following the policy π. Solving a POMDP model is equivalent to finding the optimal policy π_t^*, which achieves the maximum of the value function at any time t:

$$\pi_t^* := \arg\max_{\pi} V_t^{\pi^*}(b_t), \ \forall b_t, \forall t.$$

3.2. MPOMDP Definition

The issue of model ambiguity motivates the formulation of the MPOMDP. Suppose there are M ($M < \infty$) different POMDP models, where all models share the same state space, action space, and observation space, but may have different model parameters of initial distribution functions, transition probability and observation probability matrices, or reward functions. We assume that any of the models could be the "true" model describing the underlying stochastic system to study. However, we are unable to pick a single model because of the lack of information on the true model. The way the MPOMDP model tackles this issue is to consider all different POMDP models simultaneously by assigning a weight to the objective function of each POMDP model according to a belief vector introduced later. The model learns and updates the weights (belief vector) as the system progresses, to optimize the weighted sum of the objective functions of all POMDP models. A formal definition of the MPOMDP model is given as follows.

Definition. An MPOMDP model $\mathcal{M}$ is defined as a tuple $(\mathcal{M}_1, ..., \mathcal{M}_M, \lambda)$, where M is the number of POMDPs, each $\mathcal{M}_m = (S, b_0^m, A, P^m, O, F^m, r^m, T)$ is a POMDP model as defined in Definition 3.1 for $m = 1, ..., M$, and $\lambda = (\lambda_1, ..., \lambda_M)$ is a vector of the initial model weights for all M POMDP models such that

$$\lambda_m \in (0,1), \ \forall m = 1, ..., M, \text{ and } \sum_{m=1}^{M} \lambda_m = 1.$$

The initial weight parameter vector λ in Definition 3.2, can be viewed as a vector with elements λ_m that are the probability that the model $\mathcal{M}_m$ is the true model describing the underlying stochastic system to study at the starting time, for $m = 1, ..., M$. The initial λ vector is usually given by some prior knowledge about the relative importance or preference of each model, or set as a non-informative prior distribution. Then, every time a system output is observed, the model and state probability distributions are updated. The notion of a belief vector will arise often and can be defined as follows:

Definition. (Belief Vector) For an MPOMDP model $\mathcal{M}$, the belief vector b_t of $\mathcal{M}$ at time t is defined as

$$b_t := (b_t^1, ..., b_t^M),$$

where each element is itself a vector

$$b_m^t = (b_m^t(s_1), ..., b_m^t(s_{|S|})),$$

and each $b_t^m(s_k)$ is the probability that the underlying model of the stochastic system is model $\mathcal{M}_m$ and the system is in state s_k at time t, for $m = 1, ..., M$, $t = 1, ..., T$, and all state $s_k \in S$. Specifically, at $t = 0$, the initial belief vector of $\mathcal{M}$ is defined as

$$b_0 := (b_0^1, ..., b_0^M) \circ \lambda,$$

where $b_0^1, ..., b_0^M$ are the initial belief vectors for models $\mathcal{M}_1, ..., \mathcal{M}_M$ respectively, λ is the initial belief weight, and $\circ$ denotes the Hadamard product.

To define the optimal value problem in an MPOMDP model $\mathcal{M}$, we first describe the process flow. Initially, the underlying system is true described by one of the given POMDP models, and is in one of the states in the state space. However, the decision-maker knows neither which of the given POMDP models is the true model nor the state of the system. Instead, the decision-maker obtains an initial weight parameter λ in advance based on prior knowledge and the estimate of the initial belief vector (i.e., the probability distribution over states) in each model. Then, at the beginning of each time period, with the estimate of the belief vector of the MPOMDP model, the decision-maker can take action to influence the dynamics of the underlying system. The system then generates an output according to the chosen action, the state of the system, and the observation probability function of the actual underlying POMDP model. To select an action, the decision-maker approximates the observation probabilities by an adjusted observation probability function using the model belief, which will be discussed in detail in the next section. After observing the output, the immediate reward for each POMDP is computed according to the estimate of state distribution, the action taken, the output from the system, and its reward function. Lastly, the MPOMDP belief vector is updated. The objective of the MPOMDP is to optimize the expectation of the sum of the immediate rewards under all possible POMDPs until the end of the time horizon, according to estimated believes. Figure 1 illustrates the process flow of the optimal value problem in an MPOMDP.

We now define the optimal value problem of an MPOMDP model $\mathcal{M}$ as follows.

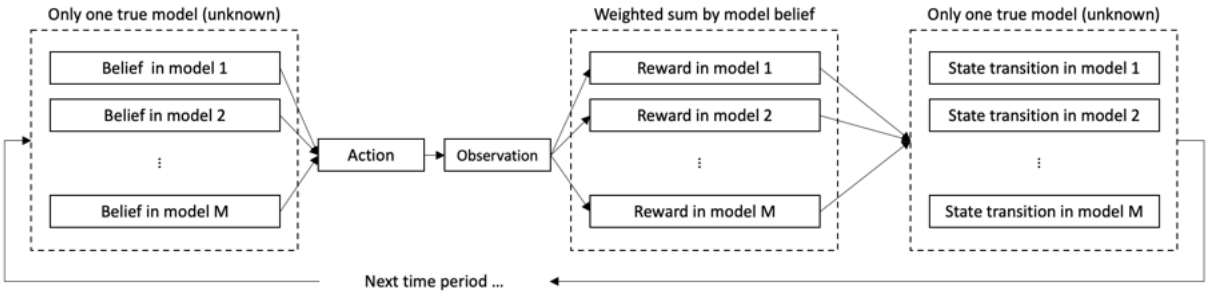

Figure 1: Illustration of the process flow of the optimal value problem in an MPOMDP.

Definition. (Optimal Value Problem) For an MPOMDP model $\mathcal{M}$, the optimal value problem entails finding the optimal policy $\pi^* = (\pi_0^*, ..., \pi_T^*)$ that achieves the maximum value function defined as follows:

$$V_t^{\pi^*}(b_t) := \max_{\pi} \sum_{m=1}^{M} V_t^{m,\pi}(b_t^m), \ \forall b, \ \forall t,$$

where $b_t = (b_t^1, ..., b_t^M)$, b_t^m is the belief vector in $\mathcal{M}_m$, and $V_t^{m,\pi}(b_t^m)$ is the value function of policy π in $\mathcal{M}_m$ defined as

$$V_t^{m,\pi}(b_t^m) := \mathbb{E}^{m,\pi}[\sum_{k=t}^{T} \gamma^{t-k} r^m(s_k, a_k, o_k) | b_t^m], \ \forall b_t^m, \forall t \leq T,$$

with the expectation taken over all possible state, action, and observation trajectories following policy π in model $\mathcal{M}_m$ for $m = 1, ..., M$.

In Definition 3.2, the optimal value problem of an MPOMDP model $\mathcal{M}$ is defined upon the initial weight parameter vector λ, which is pre-specified in the definition of $\mathcal{M}$. The initial weight parameter vector λ is integrated into the MPOMDP belief vector since time $t = 1$, as defined in Definition 3.2, so that we do not need to add duplicate weights in the value functions definition in Definition 3.2.

Starting from here, we will drop the subscript t of b_t in $V_t^{\pi}(b_t)$ when there is no confusion that V_t^{π} is the value function at time $t \leq T$. We also substitute $V_t^{\pi^*}$ by V_t^*, or even by V_t, for all $t \leq T$ as a simplification if there is no confusion.

4. Model Properties

In this section, we discuss some structural properties of the proposed MPOMDP, which show how the model addresses the issue of model ambiguity in stochastic sequential decision-making,

and motivate the solution methods introduced in the next section. We first provide the adjusted observation probability function and the belief update formula for the optimal value problem of an MPOMDP mentioned in Section 3.2. Then, we show that the optimal value problem of an MPOMDP can be reformulated as a new POMDP model with a larger state space, confirming that all properties of POMDP models will hold.

We first calculate the observation probability with respect to the belief vector in the optimal value problem. At each epoch, although the system output is generated according to the state, action, and observation probability function, the true underlying model is hidden from the decision maker. The formula below provides a method to calculate the probability of observing a certain observation, and guide decision-making.

Definition. Given an MPOMDP model $\mathcal{M}$, consider its optimal value problem defined in Definition 3.2. Then, at any time $t \geq T$, given the belief vector b, the probability of observing output o when action a is taken is

$$\begin{aligned} \mathbb{P}(o|b,a) :&= \sum_{s\in S, m=1,\ldots,M} \mathbb{P}(o,(s,m)|b,a) \\ &= \sum_{s\in S, m=1,\ldots,M} \mathbb{P}((s,m)|b,a)\mathbb{P}(o|(s,m),b,a) \\ &= \sum_{s\in S, m=1,\ldots,M} b^m(s)F^m(s,a,o), \end{aligned} \tag{1}$$

for all $o \in O$, belief vector b, and $a \in A$.

Given the observation probability of the optimal value problem, we can then show that the belief vector of the MPOMDP model is a sufficient statistic for decision-making at each time period. This property is important because it implies that the distribution over the states at each time period does not require all historical information of actions and observations.

Proposition 1. *Given an MPOMDP model $\mathcal{M}$, consider its optimal value problem defined in Definition 3.2. Then, the belief vector b_t defined in Definition 3.2 is a sufficient statistic of the past sequence of actions and observations until time t for $t = 0, 1, ..., T$.*

We can now provide the belief update formula after taking action and observing an output at each time period in an MPOMDP model.

Proposition 2. *Consider the optimal value problem of an MPOMDP model $\mathcal{M}$. Suppose $b_t = (b_t^1, ..., b_t^M)$ is the belief vector of $\mathcal{M}$ at the beginning of time t, and observation o is observed after*

taking action a, then the belief vector $b_{t+1} = (b^1_{t+1}, ..., b^M_{t+1})$ of $\mathcal{M}$ at the time $t+1$ is given by

$$\mathbb{P}((s_{t+1}, m_{t+1})|o, b, a) = \frac{\sum_{s_t} F^{m_{t+1}}(s_t, a, o) P^{m_{t+1}}(s_{t+1}, a, s_t) b^{m_{t+1}(s_t)}}{\sum_{s_{t+1}, m_{t+1}} \sum_{s_t} F^{m_{t+1}}(s_t, a, o) P^{m_{t+1}}(s_{t+1}, a, s_t) b^{m_{t+1}(s_t)}}.$$

For simplicity, we use $b_{t+1} = \Lambda(b_t|a, o)$ to denote the belief update formula given action a and observation o at time t for $t = 0, 1, ..., T-1$.

The proofs of Proposition 1 and 2 are shown in Appendix.

Remark 1. *In Proposition 2, the belief vector is updated by the Bayesian formula, which calculates a posterior distribution over models and states. Even if none of the POMDPs considered in the MPOMDP is the true model for the study object, the belief update formula in Proposition 2 still assigns a higher weight to the model with a greater probability of generating the observed outputs.*

Proposition 2 shows that the MPOMDP model is able to learn the model distribution over time from past actions and observations. Propositions 1 and 2 also show that, an MPOMDP can be viewed as a POMDP, or continuous-state MDP, when solving the optimal value problem, where the state is specified by the belief vector of the MPOMDP model. The state transition probabilities can be calculated by Proposition 2. Although the dimensionality of the state in such POMDP can be extremely high, it helps understand the structure of the MPOMDP. For example, it immediately proves the existence of a deterministic and Markovian optimal policy for the optimal value problem of an MPOMDP, as shown in the following corollary.

Corollary 1. *When considering the optimal value problem defined in Definition 3.2, an MPOMDP model can be reformulated as a POMDP model. Therefore, the MPOMDP inherits properties of POMDPs, including that there always exists an optimal policy that is deterministic and Markovian with respect to the belief vector at each time period.*

It follows from Corollary 1 that the MPOMDP model inherits the properties of the POMDP model, including that the optimal value problem is piecewise linear and convex, which will be used as the basis for the solution methods in the next section. Moreover, Corollary 1 serves as the basis for understanding the effect of model ambiguity in POMDPs. Later in Section 6, we will use computation experiments to demonstrate the effect of model ambiguity on the optimal value function and policy, and the corresponding VSS and EPVI in each example. We refer the audiences to Chapter 4.4 of Li (2021) for the proof of non-negative VSS and EPVI in MPOMDPs.

5. Solution Methods

In this section, we discuss solution methods for the proposed MPOMDP model. We start with an exact solution method, which generalizes the one-pass algorithm by Smallwood & Sondik (1973) for POMDP models. However, because of the curse of dimensionality and the curse of history, the exact solution method can take a long time to complete, even for small problems. Therefore, we introduce two approximation methods that can get near-optimal solutions efficiently. We also prove that the proposed approximation methods converge asymptotically. Finally, we compare the performance of the approximation methods in the next section.

5.1. Optimal value function and exact solution method

Consider an MPOMDP model $\mathcal{M}$, we can denote V_t as its optimal value function at any time t. Recall the recursion formula, i.e., the optimality equation, of the value function

$$V_t(b) = \max_{a \in A} \{ r(b,a) + \sum_{o \in O} \mathbb{P}(o|b,a) V_{t+1}(\Lambda(b|a,o)) \}, \ \forall b, \ \forall t,$$

with the boundary condition

$$V_T(b) = \max_{a \in A} r(b,a),$$

where

$$r(b,a) = \sum_{o \in O} \sum_{m} \sum_{s \in S} \mathbb{P}(o|b,a) r^m(s,a,o) b^m(s)$$

is the expected immediate reward, $\mathbb{P}(o|b,a)$ is the observation probability of output o, and $\Lambda(b|a,o))$ is the belief update formula provided by Proposition 2, given the current belief vector b and action a is taken for all possible beliefs b and actions a.

It follows from Corollary 1 that the optimal value function of an MPOMDP model $\mathcal{M}$ is piecewise-linear and convex in the belief vector b, and can be written as

$$V_t(b) = \max_{\alpha \in \mathcal{A}_t} \alpha \cdot b, \ \forall b,$$

where $\mathcal{A}_t$ is a set of $|S| \times M$-dimension vectors (also referred to as α-vectors) for all time periods t. Given this property, solving the optimal value problem of $\mathcal{M}$ is equivalent to finding the minimal α-vector sets $\mathcal{A}_t$ for all time periods t. This can be done by backward induction with a linear programming-based pruning algorithm for non-dominated α-vectors, which is a generalization of the solution method for POMDPs. We refer the audiences to Chapter 4.5 of Li (2021) for the detailed steps of the exact solution method for MPOMDPs. The reminder of this section will focus on the more practical approximation methods for MPOMDPs.

5.2. Sampling-based approximation methods

Although the value function is a continuous function of the belief vector, there are only a finite number of reachable belief vectors at each time period if starting from a certain initial belief at the first epoch. In other words, in order to find the optimal value function and the optimal policy of an MPOMDP model, it is sufficient to calculate the value function at all reachable belief vectors at each time period given a fixed initial belief vector. However, the number of reachable belief vectors increases exponentially in the number of possible actions, observations, and time periods. The ideal case is that we only need to know the value function at all reachable belief vectors under the optimal policy starting from the end of time horizon, and then use backward induction to calculate the optimal value function at all time periods. Unfortunately, the optimal policy can not be determined without knowing the optimal value function.

As we showed, the optimal value function is piecewise-linear and convex, and can be represented by the supremum of a set of linear functions (α-vectors). Using this property, if one can identify the dominating α-vectors at some sampled reachable belief vectors, then their supremum also gives a lower bound approximation of the optimal value function over the entire belief vector space. We leverage this fact to propose two sampling-based approximation methods for the proposed MPOMDP model. The first method uses ϵ-greedy sampling to balance exploitation and exploration of the reachable belief points based on the most recent estimate of the optimal value function. This is motivated by the ϵ-greedy algorithm for reinforcement learning problems as discussed in Sutton & Barto (2018). The second method is a tree-based branch-and-bound method, which seeks to improve the sampling efficiency of the ϵ-greedy method by branching to the belief vector where the most recent estimate has the largest error at each time period.

5.2.1. An ϵ-greedy sampling method

Denote $\mathcal{M}$ as the MPOMDP model to solve. To initialize, we sample a uniform grid of the entire space of the belief vector at each time period:

$$B_t^0 = \{0, \frac{1}{N}, \frac{2}{N}, ..., 1\}^M \subset [0,1]^M, \forall t = 1, ..., T,$$

where the superscript of B_t^0 denotes the number of iterations (includes 0), N controls the number of belief vectors and density of the uniform grid. With a finite set of grid points, an approximate backward induction works as follows. First, at the end of time horizon T, similar to Section 5.1, we calculate the set of α-vectors as

$$\mathcal{A}_T = \{(\alpha_{T,a}^1, ..., \alpha_{T,a}^M) | a \in A\},$$

where

$$\alpha_{T,a}^m(s) = \sum_{o \in O} r^m(s,a,o) F^m(s,a,o), \ \forall m, \forall a, \forall s.$$

Now, instead of keeping all α-vectors in $\mathcal{A}_T$, we only keep the ones that are non-dominated at the belief vectors in B_T^0, which gives $\hat{\mathcal{A}}_T$

$$\hat{\mathcal{A}}_T := \{\alpha \in \mathcal{A}_T | \alpha = \arg\max_{\alpha} \alpha \cdot b \text{ for some } b \in B_T^0\}.$$

Since $\hat{\mathcal{A}}_T \subset \mathcal{A}_t$, then $\hat{V}_T$ defined as

$$\hat{V}_T(b) := \max_{\alpha \in \hat{\mathcal{A}}_T} \alpha \cdot b, \ \forall b$$

gives a lower bound estimate of the optimal value function V_T at time T. Next, we go backward to time $T-1$. As described in Section 5.1, we can calculate the α-vectors at time $T-1$ using the optimality equation (1). But here, instead of using $\mathcal{A}_T$, we only use its subset $\hat{\mathcal{A}}_T$ to derive the set of α-vectors at time $T-1$, denoted as $\tilde{\mathcal{A}}_{T-1}$. It is easy to see that $\tilde{\mathcal{A}}_{T-1}$ is a subset of $\mathcal{A}_{T-1}$, which is the set of all α-vectors at time $T-1$ if using $\mathcal{A}_T$ other than $\hat{\mathcal{A}}_T$ in backward induction. Again, instead of keeping all elements in $\tilde{\mathcal{A}}_{T-1}$, we only keep the ones that are dominating at the belief vectors in B_{T-1}^0, which gives $\hat{\mathcal{A}}_{T-1}$,

$$\hat{\mathcal{A}}_{T-1} := \{\alpha \in \tilde{\mathcal{A}}_{T-1} | \alpha = \arg\max_{\alpha} \alpha \cdot b \text{ for some } b \in B_{T-1}^0\}.$$

Since $\hat{\mathcal{A}}_{T-1} \subset \tilde{\mathcal{A}}_{T-1} \subset \mathcal{A}_{T-1}$, then $\hat{V}_{T-1}$ defined as

$$\hat{V}_{T-1}(b) := \max_{\alpha \in \hat{\mathcal{A}}_{T-1}} \alpha \cdot b, \ \forall b$$

gives a lower bound estimate of the optimal value function V_{T-1} at time $T-1$. We continue backward following the steps above until time $t=0$, which gives a lower bound on the optimal value function at all time periods $\hat{V}_0, \hat{V}_1, ..., \hat{V}_T$.

The next step is to modify the grid of belief points $B_1^0, ..., B_T^0$ to improve the estimates of value functions. Starting from time $t=0$, denote b_0 as the initial belief vector. The current estimate of the value function at time $t=1$ is used to find the optimal action to take under the current optimal value function approximation at time $t=0$, which is given by

$$\hat{a} = \arg\max_a \sum_m \sum_{s \in S} b_0^m(s) \{r^m(s,a) + \sum_{o \in O} \sum_{s' \in S} \gamma F(s,a,o) P^m(s,a,s') \hat{V}_1^m(\Lambda(b_0^m | a, o))\}.$$

Notice that $\hat{a}$ may be sub-optimal, because it is selected using an approximation of the expected future value-to-go. Next, action $\hat{a}$ is selected with probability $1-\epsilon$ and a randomly sampled

alternative action with probability ϵ, for some $\epsilon \in (0,1)$, to encourage the exploration of other actions that can potentially be better than $\hat{a}$. After taking the selected action, denoted as a_0, we then randomly sample an output of the system o_0 according to the observation probability matrix F. Given the action a_0 and observation o_0, the belief vector at time $t = 1$ can be updated by

$$b_1 = \Lambda(b_0|a_0, o_0).$$

We then add b_1 into B_1^0 to get $B_1^1 = B_1^0 \cup \{b_1\}$. Now starting from belief b_1 at time $t = 1$, we repeat the steps above to sample the belief vectors $b_2, ..., b_T$ until the end of time horizon T, and get the new sets B_t^1 for $t = 2, .., T$. Collectively, the complete set of backward and forward steps is one iteration of the ϵ-greedy sampling method.

In the next iteration, we conduct the backward induction steps on the new belief vector set $B_T^1, ..., B_1^1$, and then sample the new belief vectors to get the new sets B_t^2 for $t = 1, .., T$. We repeat these iterations until a stopping criterion is satisfied. For example, if the difference between the approximate value functions in two consecutive iterations is below some threshold. This completes the steps of our proposed approximation algorithm based on ϵ-greedy sampling. We summarize the complete algorithm in Algorithm 1.

As we can see from Algorithm 1, if we denote $\bar{V}_t^i$ for all t as the lower bound estimates of the optimal value functions after the i^{th} iteration, then $\bar{V}_t^i$ is determined by the set of sampled belief vectors B_t^i, which is generated by random sampling, for each time t. Next, we show that the lower bound estimates $\bar{V}_t^i$ converge to V_t in probability at all reachable belief vectors for all time periods t, as the number of iterations i goes to infinity. The proof is shown in the Appendix.

Theorem 1. *For a given MPOMDP model $\mathcal{M}$, denote $\tilde{B}_t$ as the set of all reachable belief vectors at time $t \leq T$ starting from the initial belief vector b_0 under all possible policies for actions. Denote V_t as the optimal value function at time $t \leq T$, and $\hat{V}_t^i$ as the lower bound estimate of the optimal value function at time $t \leq T$ given by the i^{th} iteration of Algorithm 1. Then for all $t \leq T$, for any $b \in \tilde{B}_t$,*

$$\hat{V}_t^i(b) \to V_t(b) \text{ in probability, as } i \to \infty.$$

Although Theorem 1 shows that Algorithm 1 converges asymptotically to the optimal value function of the true underlying POMDP, we found through experimentation that the value function approximated at each iteration of Algorithm 1 is not monotone. In other words, the lower bound estimate of the optimal value function given by Algorithm 1 may not be monotone non-decreasing

Algorithm 1: Approximation algorithm based on ϵ-greedy sampling.

Input : MPOMDP model $\mathcal{M}$, ϵ

Output: $\hat{V}_t$

Initialize B^0 as a uniform grid and $i = 0$;

repeat

  At time T, calculate $\mathcal{A}_T$;

  $\hat{\mathcal{A}}_T = \{\alpha \in \mathcal{A}_T | \alpha = \arg\max_\alpha \alpha \cdot b \text{ for some } b \in B_T^i\}$;

  $\hat{V}_T = (b) = \max_{\alpha \in \hat{\mathcal{A}}_T} \alpha \cdot b, \ \forall b$;

  for $t = T-1, ..., 0$ **do**

    Calculate the set of α-vectors $\tilde{\mathcal{A}}_t$ at time t by backward induction using $\hat{\mathcal{A}}_{t+1}$;

    $\hat{\mathcal{A}}_t = \{\alpha \in \tilde{\mathcal{A}}_t | \alpha = \arg\max_\alpha \alpha \cdot b \text{ for some } b \in B_t^i\}$;

    $\hat{V}_t(b) = \max_{\alpha \in \hat{\mathcal{A}}_t} \alpha \cdot b, \ \forall b$;

  end

  for $t = 0, ..., T-1$ **do**

    $\hat{a} = \arg\max_a (r(a) + \sum_{o \in O} \mathbb{P}(o|b_t, a) V_{t+1}(\Lambda(\hat{b_t}|a, o)))$;

    $$a_t = \begin{cases} \hat{a}_t, & \text{with probability } 1 - \epsilon \\ \text{a random action}, & \text{with probability } \epsilon \end{cases};$$

    Sample an output o_t according to b_t and F;

    $b_{t+1} = \Lambda(b_t | a_t, o_t)$;

    $B_t^{i+1} = B_{t+1}^i \cup \{b_{t+1}\}$;

  end

  $i = i + 1$;

until *some stopping criterion*;

as we keep adding new reachable belief vectors to exploit. We use the following proposition to address this fact. The proof is by construction and given in the Appendix.

Proposition 3. *Denote $\hat{V}_t^i$ as the lower bound estimate of the optimal value function at time $t \leq T$ given by the i^{th} iteration of Algorithm 1. Then, $\hat{V}_t^i$ is not monotone non-decreasing in i. In other words, there may exists an MPOMDP model $\mathcal{M}$ such that $\exists t$, $\exists b$, $\exists i$,*

$$\hat{V}_t^{i+1}(b) - \hat{V}_t^i(b) < 0.$$

We found that such non-monotone behavior could slow the rate of convergence of Algorithm 1. We experimented with modifications of Algorithm 1 to improve the convergence rate. For example, instead of using a fixed ϵ, we tried to adaptively change the value of ϵ over iterations; we tried to sample multiple outputs and append more than one belief vector to the belief vector set in each iteration; we also tried to design a rule to remove some existing belief vectors in the belief vector set. However, we found the random sampling of system outputs is the greatest barrier to accelerating the convergence rate. Therefore, we propose another approximation algorithm that uses a branch-and-bound method to improve Algorithm 1

5.2.2. A Tree-based branch-and-bound method

Similar to the ϵ-greedy sampling method discussed above, we initially create a uniform grid of the entire space of the belief vector at all time period B_t^0 for $t = 1, ..., T$. Starting from the end of time horizon T, we first calculate $\mathcal{A}_T$ as the set of all α-vectors, and $\hat{\mathcal{A}}_T$ as the set of α-vectors that are dominating at the belief vectors in B_T^0. With $\hat{\mathcal{A}}_T$, $\hat{V}_T$ gives a lower bound on V_T. We use B_T^0 to derive an upper bound of V_T at iteration 0 as follows. For each $b \in B_T^0$, calculate $v_T(b)$ as

$$v_T(b) = \max_{\alpha \in \mathcal{A}_T} \alpha \cdot b,$$

and define $v_T(B_T^0)$ as the set

$$v_T(B_T^0) := \{(b, v_T(b)) | b \in B_T^0\}.$$

Then, since V_T is a piecewise-linear and convex function, $v_T(B_T^0)$ can be used to find an upper bound $\bar{V}_T$ of V_T by the following linear program, where for all belief vector $b \in B_T^0$,

$$\begin{aligned} \bar{V}_t(b, v_T(B_T^0)) := \min_\lambda \quad & \textstyle\sum_{b' \in B_T^0} \lambda_{b'} v_T(b') & & \\ \text{s.t.} \quad & \textstyle\sum_{b' \in B_T^0} \lambda_{b'} &=& 1, \\ & \lambda_{b'} &\geq& 0, \quad \forall b' \in B_T^0 \\ & \textstyle\sum_{b' \in B_T^0} \lambda_{b'} b' &=& b. \end{aligned}$$

Next, at time $T-1$, similar to the procedure in Section 5.2.1, we use $\hat{\mathcal{A}}_T$ and $\hat{V}_T$ to derive the lower bound estimate $\hat{\mathcal{A}}_{T-1}$ and $\hat{V}_{T-1}$. For the upper bound estimate, for each $b \in B^0_{T-1}$, calculate $u_{T-1}(b)$ as

$$u_{T-1}(b) = \arg\max_a \sum_m \sum_{s\in S} b_0^m(s)\{r^m(s,a) + \sum_{o\in O}\sum_{s'\in S} \gamma F(s,a,o)P^m(s,a,s')\bar{V}_t^m(\Lambda(b_0^m|a,o))\},$$

and define $u_{T-1}(B^0_{T-1})$ as the set

$$u_{T-1}(B^0_{T-1}) := \{(b, u_{T-1}(b))|b \in B^0_{T-1}\}.$$

Then, since V_{T-1} is piecewise-linear and convex, the solution of $\bar{V}_{T-1}(b, u_{T-1}(B^0_{T-1})))$ gives an upper bound of $V_{T-1}(b)$ for all b. We can repeat these steps for time $T-2, ..., 0$ to get the lower bound estimates $\hat{V}_{T-2}, ..., \hat{V}_0$ and upper bound estimates $\bar{V}_{T-2}, ..., \bar{V}_0$.

The next step is to modify the grid of belief vectors $B^0_1, ..., B^0_T$. Starting from time $t = 0$, denote b_0 as the initial belief vector. Similar to the ϵ-greedy sampling method of Algorithm 1, find the currently best action $\bar{a}$ given by

$$\bar{a} = \arg\max_a \sum_m \sum_{s\in S} b_0^m(s)\{r^m(s,a) + \sum_{o\in O}\sum_{s'\in S} \gamma F(s,a,o)P^m(s,a,s')\bar{V}_1^m(\Lambda(b_0^m|a,o))\},$$

and take action a_0 to be $\bar{a}$ with probability $1-\epsilon$ and a randomly sampled action with probability ϵ, for some $\epsilon \in (0,1)$. After taking action a_0, instead of randomly sampling a system output, in this case, we select o_0 as follows

$$o_0 = \arg\max_{o\in O}(\bar{V}_1(\Lambda(b_0|a_0,o)) - \hat{V}_1(\Lambda(b_0|a_0,o))).$$

In other words, we select the system output where the current estimate of the value function has the largest error, so that it needs more exploitation in the next iteration. With a_0 and o_0, we then add the updated belief vector $b_1 = \Lambda(b_0|a_0,o_0)$ into B^0_1 to get $B^0_1 \cup \{b_1\}$, and similarly get B^1_t for $t = 2, ..., T$.

In the next iteration, we repeat all the steps above to get new estimates of the lower and upper bound of the value function, and new belief sets until a stopping criterion is satisfied. The detailed steps of the branch-and-bound approximation method are given in Algorithm 2. Notice that at any node of the scenario tree, if there exists another node at the same level (observation or action node) whose lower bound value is greater than the upper bound value of this selected node, then this node can be pruned. Note that we did not put the pruning steps in Algorithm 2 for brevity. Rather, it is implicitly assumed the pruned node will not be selected in the future automatically.

Algorithm 2: The tree-based branch-and-bound approximation method.

Input : MPOMDP model $\mathcal{M}$, ϵ

Output: $\hat{V}_t$

Initialize B^0 as a uniform grid and $i = 0$;

repeat

  At time T, calculate $\mathcal{A}_T$;

  $\hat{\mathcal{A}}_T = \{\alpha \in \mathcal{A}_T | \alpha = \arg\max_\alpha \alpha \cdot b \text{ for some } b \in B_T^i\}$;

  $v_T(B_T^i) := \{(b, v_T(b)) | b \in B_T^i\}$;

  for $t = T-1, ..., 0$ **do**

    Calculate the set of α-vectors $\tilde{\mathcal{A}}_t$ at time t by backward induction using $\hat{\mathcal{A}}_t$;

    $\hat{\mathcal{A}}_t = \{\alpha \in \tilde{\mathcal{A}}_t | \alpha = \arg\max_\alpha \alpha \cdot b \text{ for some } b \in B_t^i\}$;

    $\hat{V}_t(b) = \max_{\alpha \in \hat{\mathcal{A}}_t} \alpha \cdot b, \ \forall b$;

    $u_t(B_t^i) := \{(b, u_t(b)) | b \in B_t^i\}$;

    $\bar{V}_t(b) = \bar{V}_t(b, u_t(B_t^i)), \ \forall b$;

  end

  for $t = 0, ..., T-1$ **do**

    $\bar{a} = \arg\max_a (r(a) + \sum_{o \in O} \mathbb{P}(o|b_t, a) \bar{V}_{t+1}(\Lambda(b_t|a, o)))$;

    $$a_t = \begin{cases} \bar{a}, & \text{with probability } 1 - \epsilon \\ \text{a random action}, & \text{with probability } \epsilon \end{cases};$$

    $o_t = \arg\max_{o \in O} (\bar{V}_{t+1}(\Lambda(b_t|a_t, o)) - \hat{V}_{t+1}(\Lambda(b_t|a_t, o)))$.;

    $b_{t+1} = \Lambda(b_t|a_t, o_t)$;

    $B_t^{i+1} = B_{t+1}^i \cup \{b_{t+1}\}$;

  end

  $i = i + 1$;

until *some stopping criterion*;

Algorithm 2 attempts to accelerate the convergence rate of Algorithm 1 by sampling the action with the greatest upper-bound estimate, and the observation with the largest gap between the upper-bound and lower-bound estimates. So, the asymptotic convergence of Algorithm 2 is given as a corollary of Theorem 1.

Corollary 2. *For a given MPOMDP model $\mathcal{M}$, denote $\tilde{B}_t$ as the set of all reachable belief vectors at time $t \leq T$ starting from the initial belief vector b_0 following any policies. Denote V_t as the optimal value function at time $t \leq T$, and $\hat{V}_t^i$ as the lower bound estimate of the optimal value function at time $t \leq T$ given by the i^{th} iteration of Algorithm 2. Then for all $t \leq T$, for any $b \in \tilde{B}_t$,*

$$\hat{V}_t^i(b) \to V_t(b) \text{ in probability, as } i \to \infty.$$

There are two main differences between the ϵ-greedy sampling method and the tree-based branch-and-bound method introduced in this section. First, the branch-and-bound method samples the best action based on the current upper-bound, instead of the lower-bound, estimate of the value function at each time period. This can accelerate the convergence rate because exploiting a sub-optimal action will give a smaller upper bound estimate of its value function, so that it will quickly become dominated by other actions in future steps; Second, the branch-and-bound method samples the system output at each time period according to the gap between the upper and lower bound estimates at the resulting belief vector. Thus, the algorithm tends to modify the belief space grid in areas with the biggest estimation error. However, a drawback of the branch-and-bound method is that it requires more computational effort for the upper bound estimate of the value function, which can be problematic when the number of sampled belief vectors becomes large. In practice, we use the branch-and-bound method to get a warm start, and then switch to the ϵ-greedy sampling method.

6. Computational Experiments

In this section, we describe two computational experiments to illustrate the application of the proposed MPOMDP. The first experiment is a toy example with two POMDPs, which have two states, two observations, and two actions. We use this toy example to visualize the value function and optimal policy of the MPOMDP model. We also show the VSS and the EVPI in this context. Furthermore, we use the toy example to compare the performance of the two approximation methods introduced in Section 5. The second computational experiment is a case study in prostate cancer AS

based on the POMDP models studied in Li et al. (2023). We use the proposed MPOMDP to find the optimal timing of biopsies in AS when the cancer progression rate and test accuracy are assumed to be uncertain because of the existence of multiple plausible selections of model parameters.

6.1. A two-model toy example

Suppose there are two POMDP models denoted as $\mathcal{M}_m = (S, b_0, A, P^m, O, F^m, r^m, T)$ for $m = 1, 2$, which have the same state space, observation space, and action space

$$S = \{s_1, s_2\},\ O = \{o_1, o_2\},\ A = \{a_1, a_2\}$$

but the different transition and observation probabilities

$$P^1(a_1) = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix} F^1(a_1) = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix},$$

$$P^1(a_2) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} F^1(a_2) = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix},$$

$$P^2(a_1) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} F^2(a_1) = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix},$$

$$P^2(a_2) = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix} F^2(a_2) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

and the reward function

$$a_1 : r(s_1, a_1, o_1) = 2, r(s_1, a_1, o_2) = 0, r(s_2, a_1, o_1) = 0, r(s_2, a_1, o_2) = 1$$

$$a_2 : r(s_1, a_2, o_1) = 1, r(s_1, a_2, o_2) = 0, r(s_2, a_2, o_1) = 0, r(s_2, a_2, o_2) = 2.$$

with the time horizon $t = 0, 1, 2, 3, 4, 5$.

We first solve the MPOMDP model using the exact solution method, and plot the exact value function. Figure 2 shows the value function $V_0(b)$ at time $t = 0$. Notice that the argument of the value function, which is the belief vector of the MPOMDP model, is a 4-dimension vector with three degrees of freedom. Thus, we plot $V_0(b)$ for various choices of $b^2(s_1)$ to illustrate the 4-dimension function $V_0(b)$. As we can see from Figure 2, $V_0(b)$ is a piecewise linear and convex function in b. When the belief vector lies in the dark region, then the optimal action to take at time $t = 0$ will be a_1; otherwise, if the belief vector lies in the light region, then the optimal action will be a_2.

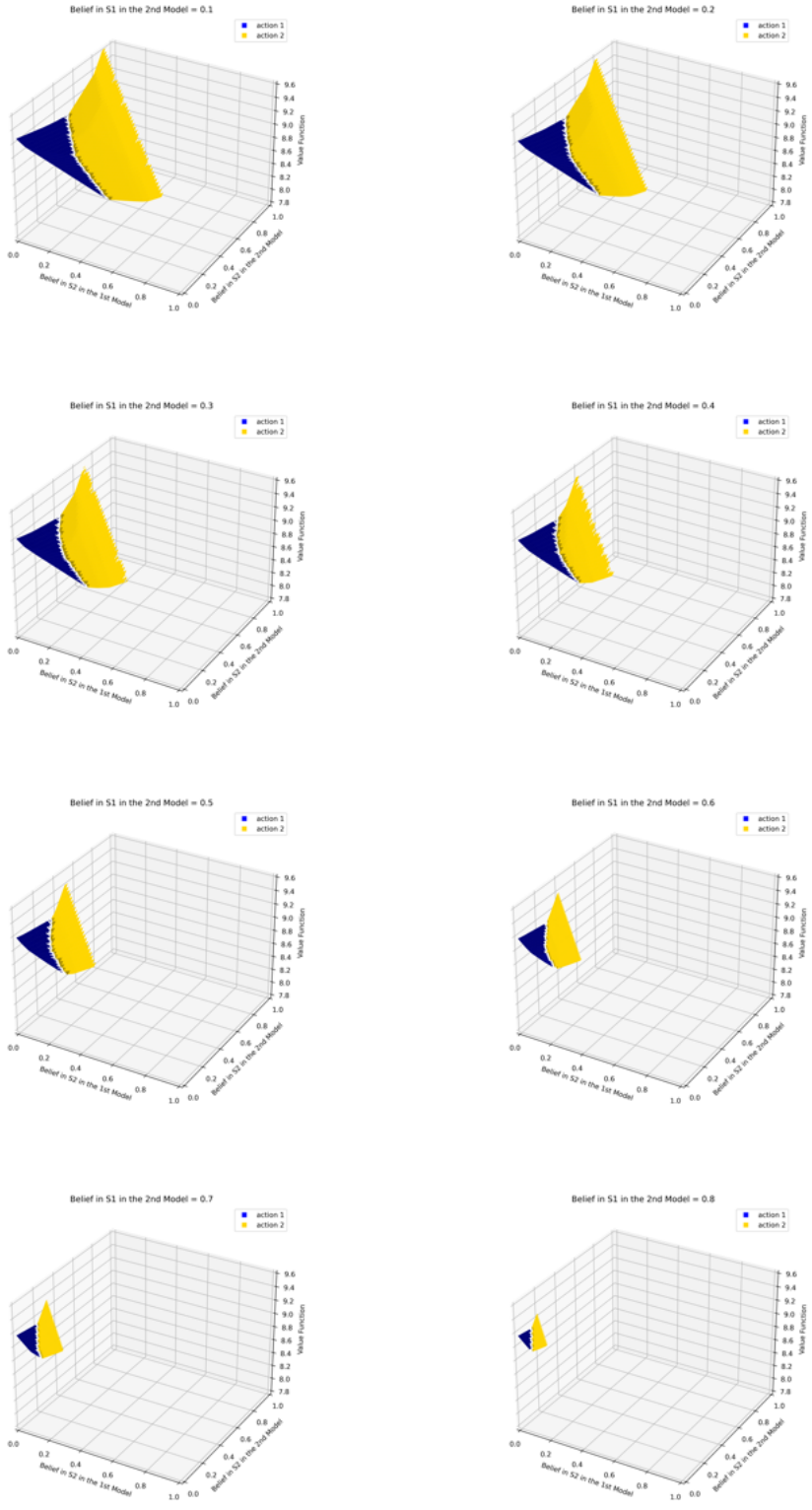

Figure 2: The value function $V_0(b)$ at time $t = 0$ for various choices of $b^2(s_1)$. $V_0(b)$ is a piecewise linear and convex function in b. When the belief vector lies in the dark region, then the optimal action to take at time $t = 0$ will be a_1; otherwise, the optimal action will be a_2

Belief vector b_0	Value of the optimal policy (Regret %)				
	Model $\mathcal{M}_1$	Model $\mathcal{M}_2$	Mean-value model	MPOMDP model	Perfect info.
(0.45, 0.05, 0.45, 0.05)	8.43 (9.52%)	7.81 (16.11%)	8.89 (4.55%)	9.08 (2.48%)	9.31 (0)
(0.45, 0.05, 0.25, 0.25)	8.10 (14.04%)	7.93 (15.82%)	9.03 (4.11%)	9.05 (3.95%)	9.42 (0)
(0.45, 0.05, 0.05, 0.45)	8.73 (8.40%)	8.03 (15.74%)	9.00 (5.54%)	9.18 (3.66%)	9.53 (0)
(0.25, 0.25, 0.45, 0.05)	8.29 (9.57%)	8.37 (8.65%)	8.24 (10.10%)	8.88 (3.18%)	9.17 (0)
(0.25, 0.25, 0.25, 0.25)	7.95 (14.28%)	8.46 (8.74%)	8.40 (9.42%)	8.95 (3.48%)	9.27 (0)
(0.25, 0.25, 0.05, 0.45)	8.60 (8.25%)	8.57 (8.60%)	8.51 (9.24%)	9.06 (3.38%)	9.38 (0)
(0.05, 0.45, 0.45, 0.05)	8.30 (9.65%)	8.86 (3.53%)	8.78 (4.40%)	9.07 (1.22%)	9.18 (0)
(0.05, 0.45, 0.25, 0.25)	7.96 (14.32%)	8.97 (3.41%)	8.92 (3.96%)	9.16 (1.41%)	9.29 (0)
(0.05, 0.45, 0.05, 0.45)	8.61 (8.39%)	9.12 (2.96%)	9.04 (3.80%)	9.27 (1.39%)	9.40 (0)

Table 1: The value function V_0 and the regrets at different initial belief vectors when applying different policies. The optimal policy of the MPOMDP model $\mathcal{M}$ dominates the optimal policies of model $\mathcal{M}_1$, model $\mathcal{M}_2$, and the mean-value model.

Next, we calculate the value of the VSS achieved by the MPOMDP model, and the EVPI. To begin with, we run a simulation study on a group of 10,000 samples where 50% of them are from model $\mathcal{M}_1$, and the other 50% are from model $\mathcal{M}_2$. We apply four different policies to the study group: (1) the optimal policy given by the POMDP model $\mathcal{M}_1$; (2) the optimal policy given by the POMDP model $\mathcal{M}_2$; (3) the optimal policy given by the mean-value POMDP model (i.e., the POMDP model with the model parameter being the mean parameter of $\mathcal{M}_1$ and $\mathcal{M}_2$); (4) the optimal policy given by the MPOMDP model $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \lambda = 0.5)$. We also compare the results with the case where we have the perfect information, and apply the optimal policy of $\mathcal{M}_1$ to patients from $\mathcal{M}_1$ and the optimal policy of $\mathcal{M}_2$ to patients from $\mathcal{M}_2$.

Table 1 shows the values of V_0 at different initial belief vectors when applying different policies, and their regrets compared to the value function given by the optimal policy with perfect information. As we can see from Table 1, the optimal policy of the MPOMDP model $\mathcal{M}$ dominates the optimal policies of model $\mathcal{M}_1$, model $\mathcal{M}_2$, and the mean-value model. This says that when model ambiguity exists, the MPOMDP model provides a better solution than ignoring the model ambiguity or averaging the model parameters.

Table 2 shows the VSS achieved by the MPOMDP and the EVPI for different initial belief vectors. For each initial belief vector, the VSS of the MPOMDP is calculated as the (relative) difference between the values of the mean-value POMDP model and the MPOMDP model; and

Belief vector b_0	$\lambda = (0.25, 0.75)$		$\lambda = (0.5, 0.5)$		$\lambda = (0.75, 0.25)$	
	VSS (%)	EVPI (%)	VSS (%)	EVPI (%)	VSS (%)	EVPI (%)
(0.45, 0.05, 0.45, 0.05)	0.11 (1.18%)	0.25 (2.71%)	0.19 (2.17%)	0.23 (2.55%)	0.58 (6.81%)	0.22 (2.39%)
(0.45, 0.05, 0.25, 0.25)	0.22 (2.41%)	0.24 (2.56%)	0.02 (0.17%)	0.37 (4.11%)	0.56 (6.65%)	0.32 (3.54%)
(0.45, 0.05, 0.05, 0.45)	0.43 (4.80%)	0.22 (2.29%)	0.18 (1.98%)	0.35 (3.80%)	0.34 (3.89%)	0.41 (4.53%)
(0.25, 0.25, 0.45, 0.05)	0.21 (2.40%)	0.38 (4.27%)	0.63 (7.70%)	0.29 (3.28%)	0.54 (6.45%)	0.20 (2.29%)
(0.25, 0.25, 0.25, 0.25)	0.05 (0.59%)	0.46 (5.18%)	0.55 (6.55%)	0.32 (3.61%)	0.69 (8.30%)	0.18 (2.03%)
(0.25, 0.25, 0.05, 0.45)	0.05 (0.50%)	0.46 (5.11%)	0.55 (6.45%)	0.32 (3.50%)	0.60 (7.10%)	0.12 (1.36%)
(0.05, 0.45, 0.45, 0.05)	0.13 (1.47%)	0.19 (2.10%)	0.29 (3.32%)	0.11 (1.23%)	0.73 (8.77%)	0.03 (0.36%)
(0.05, 0.45, 0.25, 0.25)	0.04 (0.47%)	0.22 (2.39%)	0.24 (2.66%)	0.13 (1.43%)	0.81 (9.64%)	0.00 (0.01%)
(0.05, 0.45, 0.05, 0.45)	0.10 (1.04%)	0.22 (2.34%)	0.23 (2.51%)	0.13 (1.41%)	0.73 (8.64%)	0.05 (0.59%)

Table 2: The VSS achieved by the MPOMDP and the EVPI for different initial belief vectors in the two-model example. The VSS and EVPI are more significant when the decision-maker is less certain about the model and state distribution.

the EVPI is calculated as the (relative) difference between the values of the MPOMDP model and model with perfect information. As we can see from Table 2, in general, the VSS and EVPI are more significant when the decision-maker is less certain about the model and state distribution. Table 3 also shows the percentage of true optimal action over time compared to the optimal policy when having the perfect information starting from different initial belief vectors.

Lastly, we compare the performance of the two approximation methods introduced in Section 5. We implement each approximation method with 100 iterations. Figure 3 reports the average error of V_0 in 20 runs, each with 100 iterations. As we can see from Figure 3, both methods make good progress over a small number of iterations. However, the tree-based sampling method converges more quickly with respect to the number of iterations. This is because, as discussed in Section 5, while both methods exploit the optimal action at each time period based on the current estimate of the value function, the tree-based sampling algorithm additionally calculates an upper bound estimate of the function to explore the scenarios where the current estimate has the largest error. This likely helps ensure more efficient exploration steps, and results in a faster overall convergence rate with respect to the number of iterations. On the other hand, in Table 4 we illustrate the computation time for each method on an Inter Core i7 2.6 GHz processor with 16 GB RAM. As we can see from Table 4, the tree-based sampling method takes more computation time for each run than the ϵ-greedy sampling method. Thus, although the more judicious choice of belief grid

Belief vector b_0	% of true optimal action over time				
	Model $\mathcal{M}_1$	Model $\mathcal{M}_2$	Mean-value model	MPOMDP model	Perfect info.
(0.45, 0.05, 0.45, 0.05)	59.70%	59.84%	79.33%	89.09%	100%
(0.45, 0.05, 0.25, 0.25)	57.91%	60.07%	85.38%	88.99%	100%
(0.45, 0.05, 0.05, 0.45)	72.79%	59.86%	86.31%	88.94%	100%
(0.25, 0.25, 0.45, 0.05)	59.40%	70.37%	61.50%	88.74%	100%
(0.25, 0.25, 0.25, 0.25)	57.60%	70.20%	67.76%	88.61%	100%
(0.25, 0.25, 0.05, 0.45)	72.51%	70.06%	69.18%	88.64%	100%
(0.05, 0.45, 0.45, 0.05)	60.14%	87.40%	79.42%	89.57%	100%
(0.05, 0.45, 0.25, 0.25)	58.76%	87.50%	85.45%	89.60%	100%
(0.05, 0.45, 0.05, 0.45)	73.70%	87.35%	86.81%	89.85%	100%

Table 3: The percentage of true optimal action over time compared to the optimal policy with the perfect information starting from different initial belief vectors for different policies. The optimal policy of the MPOMDP model $\mathcal{M}$ dominates the optimal policies of model $\mathcal{M}_1$, model $\mathcal{M}_2$, and the mean-value model.

	Exact method	ϵ-greedy sampling	Tree-based B&B
Mean time for each run	6836s	109s	551s
Number of α-vectors at $t = 0$	109	15	27

Table 4: Comparisons of the computational time and number of iterations of Algorithm 1 and 2 for the toy example of two-model POMDP.

modifications leads to fewer iterations for Algorithm 2, the shorter computation time per iteration of Algorithm 1 results in greater overall computation time efficiency, albeit with slightly higher error.

6.2. Case study: prostate cancer AS optimization

We implement the proposed MPOMDP model for optimizing AS for prostate cancer with imperfect information based on the POMDP models in Li et al. (2023). Prostate cancer is the most common cancer in men globally. Patients with low-risk variants of prostate cancer are recommended to join the AS, which monitors the patients by medical tests until there is a sign of progression to a high-risk variant of cancer, to avoid unnecessary treatments. The two most common medical tests in AS are the PSA test and biopsy. The PSA test is a simple blood test with almost no direct harm to patients. Biopsy is a much more accurate diagnostic test, which samples the tissue with

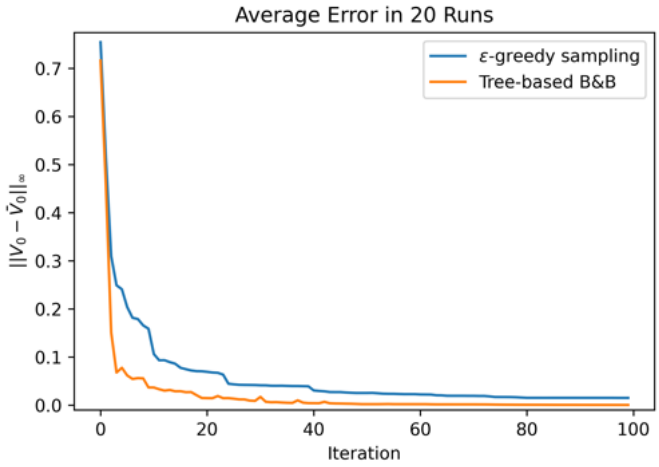

Figure 3: Comparisons of Algorithm 1 and 2 for the toy example of two-model POMDP.

hollow-core needles to examine the severity of prostate cancer; however, biopsy is still imperfect, with potential false-negative results caused by miss-sampling. Moreover, biopsy is very painful and harmful to patients. Thus, it is critical to decide the optimal timings for biopsies for each patient in prostate cancer AS.

Li et al. (2023) used a finite-horizon two-state POMDP model to optimize the biopsy policy in each of four major prostate cancer AS study centers, which include the JH hospital, the UCSF medical center, the U of T medical center, and the PRIAS project. The objective of that study was to minimize the expected delay in the detection of high-risk prostate cancer and the expected number of lifetime biopsies. The result showed that, as different patient cohorts have heterogeneous cancer progression rates and test accuracy (model parameters), the optimal biopsy policies could be quite different in the different study centers.

Our study in this paper considers the case where the model parameters are not known with certainty, and we seek a single biopsy policy that works well for models based on all four study centers. Examples may include optimizing the biopsy policy for a new patient who comes from a different area with an uncertain cancer progression rate, and for a newly initiated prostate cancer AS study that is unable to estimate the cancer dynamics (model parameters) because of a lack of data samples. For such new studies, a common strategy is to use the result from one of the previous studies as an approximate solution. The proposed MPOMDP model in this paper allows the decision-maker to trade off all previous major studies instead of picking only one study ambitiously. The objective of the MPOMDP model is to minimize a weighted sum of the expected delays in the

detection of high-risk prostate cancer and the expected number of lifetime biopsies in four study centers.

We first describe the MPOMDP model formulation $\mathcal{M}$ for optimizing prostate cancer AS. The decision epochs here are discrete annual time periods until age 75, which is the recommended stopping time for AS with the consideration of other cause mortality rates. There are two states in S, which are low-risk cancer state (LR) and high-risk cancer state (HR). The set A contains two actions that are "defer biopsy" and "conduct biopsy". At each decision epoch after taking action, there will be observations of PSA test and biopsy (if conducted). For the PSA test, we divide all possible outcomes into three intervals: $I_1 = [0, 4]$, $I_2 = (4, 10]$, and $I_3 = (10, \infty)$ (ng/mL); For biopsy, the possible outcomes are negative, positive, or null (no biopsy conducted). The transition and observation probabilities in the four different study centers were estimated in Li et al. (2020) and listed in Tables 5 and 6 for convenience. In Table 5, the misclassification error at diagnosis denotes the initial distribution b_0, the annual cancer progression rate denotes the transition probabilities, and the biopsy sensitivity denotes the observation probabilities for the biopsy. Table 6 denotes the observation probabilities for the PSA test. Lastly, the reward function $r(s, a, o)$ is defined as

$$r(s,a,o) = \begin{cases} 0, & a = \text{Defer Biopsy},\ s = \text{LR}; \\ \theta, & a = \text{Defer Biopsy},\ s = \text{HR}; \\ \eta, & a = \text{Conduct Biopsy},\ s = \text{LR},\ o = \text{Negative}; \\ \eta, & a = \text{Conduct Biopsy},\ s = \text{HR},\ o = \text{Positive}; \\ \theta + \eta, & a = \text{Conduct Biopsy},\ s = \text{HR},\ o = \text{Negative}; \\ \text{Not Defined}, & \text{otherwise}, \end{cases}$$

where θ and η are non-negative scalars that denote the cost of one-year delayed detection of high-risk cancer and the burden of a biopsy, respectively. We set $\theta + \eta = 1$, so that varying θ and η allows computing the optimal policy for different patient preferences for the two criteria. Here we choose $\theta = \eta = 0.5$ by way of example while the weighting in practice depends on patient preferences.

Now, suppose that for a group of new patients, the decision-maker has no information about which model best describes the new patients. Traditionally, the decision-maker picks a single model based on their personal judgment/opinion about which is the best, and applies its optimal policy to new patients in practice. Here, our proposed MPOMDP model provides another solution to this problem. To show the benefit of the MPOMDP model, for each AS study, we compare the result of five different biopsy policies, which includes the four policies given by solving the JH, UCSF, U

Center	Misclassification Error at Diagnosis: b_0(LR)	Annual Cancer Progression rate: p	Biopsy Sensitivity: $(1-\gamma)$
JH	0.0583	0.0691	0.7184
UCSF	0.0809	0.1217	0.7431
U of T	0.0774	0.1016	0.7949
PRIAS	0.0653	0.0841	0.7614

Table 5: The AS-POMDP model parameters in four study centers. Abbreviations: JH, Johns-Hopkins; UCSF, University of California-San Francisco; U of T, University of Toronto; PRIAS, Prostate Cancer Research International Active Surveillance.

Center	Probability Mass Function of PSA (ng/mL): q			
	Cancer State	$I_1 = [0, 4]$	$I_2 = (4, 10]$	$I_3 = (10, \infty)$
JH	LR Cancer	0.3552	0.4311	0.2137
	HR Cancer	0.2868	0.4706	0.2426
UCSF	LR Cancer	0.0768	0.5680	0.3552
	HR Cancer	0.0678	0.5736	0.3586
U of T	LR Cancer	0.4573	0.3422	0.2005
	HR Cancer	0.3312	0.2368	0.4320
PRIAS	LR Cancer	0.1361	0.5357	0.3282
	HR Cancer	0.1094	0.5501	0.3405

Table 6: The probability mass functions of PSA in four study centers. Abbreviations: JH, Johns-Hopkins; UCSF, University of California-San Francisco; U of T, University of Toronto; PRIAS, Prostate Cancer Research International Active Surveillance; LR, low-risk; HR, high-risk.

Center	Minimum cost of the optimal policy (regret %)				
	JH model	UCSF model	U of T model	PRIAS model	MPOMDP model
JH	2.74 (0)	2.92 (6.50%)	3.84 (40.42%)	3.01 (9.89%)	2.87 (4.80%)
UCSF	2.54 (5.35%)	2.41 (0)	2.95 (22.45%)	2.68 (11.33%)	2.49 (3.33%)
U of T	2.65 (12.34%)	2.42 (2.39%)	2.36 (0)	2.77 (17.54%)	2.40 (1.72%)
PRIAS	2.59 (4.19%)	2.63 (5.54%)	3.11 (24.71%)	2.49 (0)	2.54 (2.03%)

Table 7: The optimal value (minimum cost) function in different AS studies when applying different policies.

of T, and the PRIAS POMDP models, and the policy given by solving the MPOMDP model. For the MPOMDP model, we set a non-informative initial model weight $\lambda = (0.25, 0.25, 0.25, 0.25)$.

Table 7 shows the optimal value (minimum cost) function and the regret of each biopsy policy in each AS study center. The regret is calculated as the relative difference between the chosen policy and the best policy in each study center. As we can see from Table 7, the best policy in each study center is always the optimal policy given by the corresponding POMDP model. Moreover, the optimal policy given by the MPOMDP model is always better than policies from an inconsistent POMDP model in all four study centers. For each study center, the difference between the cost of the optimal policy given by the MPOMDP and a "wrong" model (different from the study center) is the VSS achieved by the MPOMDP model; and the difference between the MPOMDP and the "right" model is the EVPI. Figure 4 shows the comparison of the mean number of biopsies and average late detection time by biopsy in years in different AS studies when applying different policies in different models. Depending on how the decision-maker trades off between the mean number of biopsies and average late detection time by biopsy, the optimal policy given by the MPOMDP model is always the closest to the true optimal policy in each study center, compared with the policy given by a wrong POMDP model.

7. Conclusion

In this paper, we introduced a new MPOMDP model to address the issue of model ambiguity in POMDP models. Motivated by the prostate cancer AS optimization problem, when there are multiple credible optimization models with the same structure but different model parameters, the proposed MPOMDP model can learn the model credibility from the system outputs over time, and seek a single optimal policy that maximizes the expected future rewards across models. We

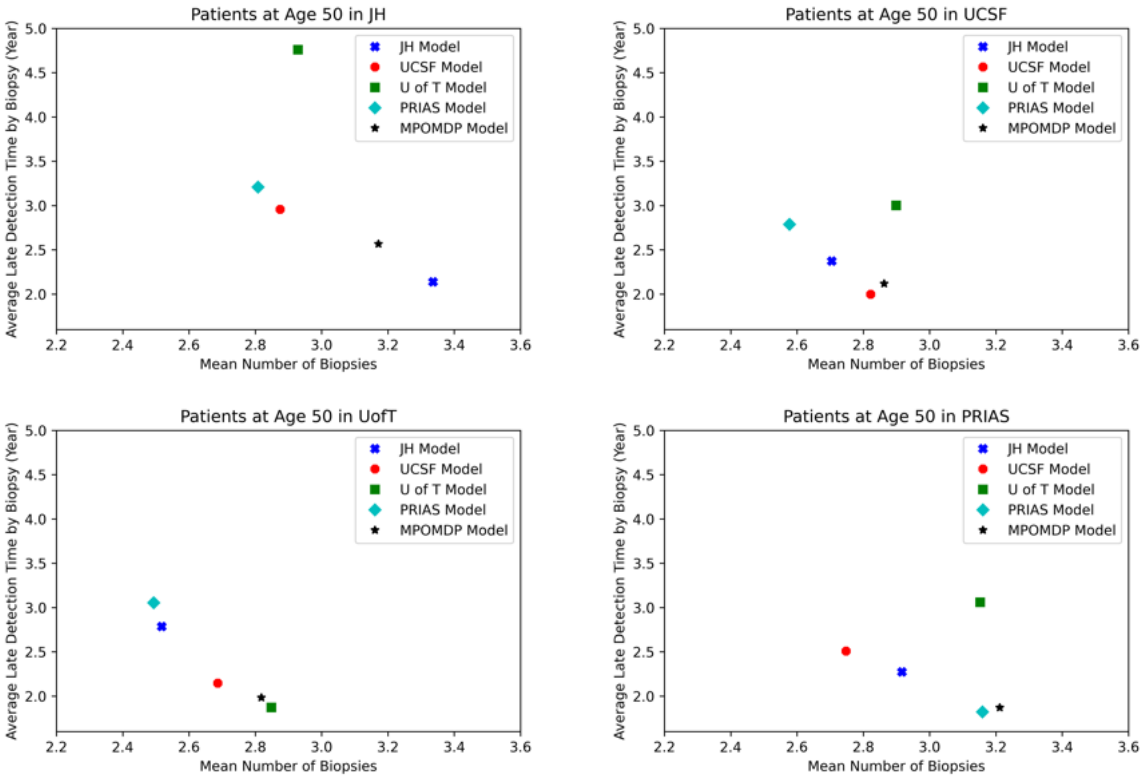

Figure 4: Comparisons of the mean number of biopsies and average late detection time by biopsy in years in different AS studies when applying the optimal policies in different models. The reward parameter is set to be $\theta = \eta = 0.5$.

also discussed some structural properties of the proposed MPOMDP model, which not only reveal the benefit of the MPOMDP model by accounting for model ambiguity, but also motivate the solution methods to MPOMDP models. We then introduced an exact solution method and two fast approximation methods to MPOMDP models, which were shown to converge asymptotically, and compared their performance in a computational experiment. Lastly, we used the example of prostate cancer AS as a case study to demonstrate how the MPOMDP model can be applied to a real-world problem to improve medical decision-making.

When applying the MPOMDP model to real-world problems, the model weight can be initialized by some prior knowledge or as a non-informative prior distribution over different POMDPs. Then, every time when there is new output from the system, the MPOMDP model can update the model belief so that more credible models will be assigned higher model weights. Notice that since the model weight is updated by the Bayesian formula, even if none of the POMDPs considered in the MPOMDP is the true model that describes a patient, the MPOMDP is still able to assign a higher weight to the model with a higher probability of generating the observed outputs. We then showed that an MPOMDP could be reformulated as a new POMDP model with an extended state space, where a state in the new POMDP model is a combination of the current model and the state in the original POMDP model. Utilizing this property, we then derived the belief update formula for both the system state and model in an MPOMDP. Further, motivated by the one-pass algorithm for POMDP models, we introduced an exact solution method to the proposed MPOMDP model. However, because of the complexity of an MPOMDP model, even for moderate-size problems, the exact solution method is not feasible in a reasonable amount of time. We then introduced two fast approximation algorithms applying the ϵ-greedy and branch-and-bound sampling methods. Thus, instead of calculating the optimal value function of the MPOMDP over the entire belief space, we only evaluate the optimal value function on a subset of reachable belief points by sampling, and then approximate the value function on other places using the samples.

Compared with the robust optimization approach, whose objective is to optimize the worst-case performance, there are three main advantages of our MPOMDP model. First, in our MPOMDP model formulation, when considering the optimal value problem, there are some nice properties, including that the belief vector is sufficient for the past information and the existence of a deterministic and Markovian optimal policy, which do not hold for robust optimization models. These properties are important because they help develop efficient solution methods so that the MPOMDP can be applicable to large real-world problems. Second, the MPOMDP model is able to learn the model

credibility for each individual over time from past actions and observations, which is not the focus in robust optimization models. Third, the MPOMDP model that optimizes a weighted-average value function by the model belief usually results in a less conservative policy than the robust optimization models that optimize the worst-case value function. On average, the MPOMDP achieves better performance than the robust optimization models.

In the computational experiments, we first used a toy example with two POMDPs to illustrate the use case of the proposed MPOMDP model. We formulated the MPOMDP with two POMDPs, solved for the optimal value function and policy exactly, and compared its performance with other traditional solutions. The results showed that the MPOMDP policy dominated the solution obtained by arbitrarily picking one POMDP model when the wrong model was selected, and the mean-value POMDP model. This was because the MPOMDP can consider the performance of both POMDPs according to the model weight learned from system outputs. We also used this example to compare the performance of two much faster approximation methods. The ϵ-greedy sampling method updated the lower bound estimate of the optimal value function in each iteration, which converged asymptotically over time. Compared with the ϵ-greedy sampling method, the branch-and-bound sampling method converged faster by maintaining an upper bound estimate of the value function. But it also required extra computational effort to calculate the upper bound estimates.

We further investigated the benefit of the MPOMDP model in a real case study of prostate cancer AS. We showed that for a new patient starting prostate cancer AS, who may be best described by one of the models in the JH, UCSF, U of T, and PRIAS study centers, the MPOMDP model found a single optimal biopsy policy that is only slightly worse than the optimal biopsy policy given by the POMDP model of the true study center, but much better than the policies given by a wrong POMDP model and the mean-value POMDP model. Given the trade-off between the biopsy burden and late detection of a cancer progression by the decision-maker, the MPOMDP model achieved the minimum expected future costs when the true model was not known with certainty. Thus, the MPOMDP model appears to offer a robust policy that protects against uncertainty when the correct model is not known with certainty. For example, Table 7 shows the regret for the MPOMDP van be substantially larger than single model POMDP policies with regrets ranging from 0 to 40%.

There are also some limitations of our work in this paper, and opportunities for future research in model ambiguity in POMDP models. First, we only focused on the optimal value problem of an MPOMDP model in this paper, where the objective was to maximize a weighted average of the value functions across different POMDPs according to the model-state belief vector. There

could be other objective functions, for example, maximizing the worst-case reward, minimizing the conditional value-at-risk, and other probability measures that are widely used in stochastic programming and robust optimization problems. However, the potential issue for considering other objective functions is the existence of an optimal policy with a simple structure, for example, a deterministic and Markovian policy, that is practical for real-world problems. We leave the theoretical and methodological study of the extension to other objective functions to future research. Second, the proposed MPOMDP model only considers a discrete finite set of models. This is different from prior stochastic optimization work, where the uncertainty sets of model parameters are continuous. However, the work of MPOMDP in this paper was motivated by the real-world application in prostate cancer AS, where there are a finite set of competing well-established models. The framework we proposed can provide a valuable foundation for studying related problems that arise in other contexts.

Acknowledgments

This work is supported in part by the National Science Foundation through Grant Number CMMI 0844511. 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. This work was also supported by the Movember Foundation. The funders did not play any role in the study design, collection, analysis or interpretation of data, or in the editing of this thesis.

References

Åström, K. J. (1965). Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, *10*, 174–205.

Ayer, T., Alagoz, O., & Stout, N. K. (2012). Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research*, *60*, 1019–1034.

Birge, J. R. (1982). The value of the stochastic solution in stochastic linear programs with fixed recourse. *Mathematical programming*, *24*, 314–325.

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

Cassandra, A. R. (1998). A survey of pomdp applications. In *Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes*. volume 1724.

Cassandra, A. R., Kaelbling, L. P., & Kurien, J. A. (1996). Acting under uncertainty: Discrete bayesian models for mobile-robot navigation. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS'96* (pp. 963–972). IEEE volume 2.

Delage, E., & Iancu, D. A. (2015). Robust multistage decision making. In *The operations research revolution* (pp. 20–46). INFORMS.

Delage, E., & Mannor, S. (2010). Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, *58*, 203–213.

Drake, A. W. (1962). *Observation of a Markov process through a noisy channel*. Ph.D. thesis Massachusetts Institute of Technology.

Erenay, F. S., Alagoz, O., & Said, A. (2014). Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing & Service Operations Management*, *16*, 381–400.

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, *30*, 257–280.

Li, W. (2021). *Data-Driven Optimization for Individualized Medical Decision-Making in Cancer*. Ph.D. thesis.

Li, W., Denton, B. T., & Morgan, T. M. (2023). Optimizing active surveillance for prostate cancer using partially observable markov decision processes. *European Journal of Operational Research*, *305*, 386–399.

Li, W., Denton, B. T., Nieboer, D., Carroll, P. R., Roobol, M. J., Morgan, T. M., & Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance consortium (2020). Comparison of biopsy under-sampling and annual progression using hidden markov models to learn from prostate cancer active surveillance studies. *Cancer Medicine*, .

Mannor, S., Mebel, O., & Xu, H. (2016). Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, *41*, 1484–1509.

Nakao, H., Jiang, R., & Shen, S. (2021). Distributionally robust partially observable markov decision process with moment-based ambiguity. *SIAM Journal on Optimization*, *31*, 461–488.

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

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rabiner, L., & Juang, B. (1986). An introduction to hidden markov models. *ieee assp magazine*, *3*, 4–16.

Ross, S. M. (1971). Quality control under markovian deterioration. *Management Science*, *17*, 587–596.

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

Schlaifer, R., & Raiffa, H. (1961). *Applied statistical decision theory*.

Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable markov processes over a finite horizon. *Operations research*, *21*, 1071–1088.

Steimle, L. N., Kaufman, D. L., & Denton, B. T. (2021). Multi-model markov decision processes. *IISE Transactions*, (pp. 1–39).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

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

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

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., & Inman, B. A. (2012). Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management*, *14*, 529–547.

Appendix A. Proofs

Proof of Proposition 1

Denote $I(t)$ as the total information available, i.e., historical actions and observations, at the end of time period t:

$$I(1) = \{a_1, o_1\}, I(t+1) = I(t) \cup \{a_{t+1}, o_{t+1}\}, \forall t \geq 1.$$

We are going to show that $b^m_{t+1}(s^m_{t+1})$ depends on $I(t)$ only through b_t for all $t \geq 1$, $s^m_{t+1} \in S$, and $m = 1, ..., M$:

$$\begin{aligned}
&b^m_{t+1}(s^m_{t+1}) \\
&= \mathbb{P}((s_{t+1}, m_{t+1})|a_t, o_t, I(t)) \\
&= \frac{\mathbb{P}((s_{t+1}, m_{t+1}), o_t|a_t, I(t))}{\mathbb{P}(o_t|a_t, I(t))} \\
&= \frac{\sum_{s_t \in S} \sum_{m_t} \mathbb{P}((s_{t+1}, m_{t+1}), (s_t, m_t), o_t|a_t, I(t))}{\mathbb{P}(o_t|a_t, I(t))} \\
&= \frac{\sum_{s_t \in S} \sum_{m_t} \mathbb{P}(o_t|(s_t, m_t), a_t, I(t))\mathbb{P}((s_{t+1}, m_{t+1})|(s_t, m_t), a_t, I(t))\mathbb{P}((s_t, m_t)|a_t, I(t))}{\mathbb{P}(o_t|a_t, I(t))} \\
&= \frac{\sum_{s_t \in S} \sum_{m_t} F^{m_t}(o, a, s_t) P^{m_t}(s_t, a_t, s_{t+1}) b^m_t(s_t)}{\mathbb{P}(o_t|a_t, I(t))}.
\end{aligned}$$

Now, we can see the numerator of $b^m_{t+1}(s^m_{t+1})$ depends on $I(t)$ only through b_t, and the denominator is just the numerator summed over all possible values of s^m_{t+1}. Thus, b_t is a sufficient statistics of $I(t)$ for all $t = 1, ..., T$. $\square$

Proof of Proposition 2

First of all,

$$\mathbb{P}((s_{t+1}, m_{t+1})|o, b, a) = \frac{\mathbb{P}((s_{t+1}, m_{t+1}), o|b, a)}{\mathbb{P}(o|b, a)}.$$

For the numerator,

$$\begin{aligned}
&\mathbb{P}((s_{t+1}, m_{t+1}), o|b, a) \\
&= \sum_{s_t} \sum_{m_t} \mathbb{P}((s_{t+1}, m_{t+1}), o, (s_t, m_t)|b, a) \\
&= \sum_{s_t} \sum_{m_t} \mathbb{P}(o|(s_{t+1}, m_{t+1}), (s_t, m_t), b, a)\mathbb{P}(s_{t+1}, m_{t+1}), (s_t, m_t)|b, a) \\
&= \sum_{s_t} \sum_{m_t} \mathbb{P}(o|(s_t, m_t), a)\mathbb{P}(s_{t+1}, m_{t+1})|(s_t, m_t), a)\mathbb{P}((s_t, m_t)|b).
\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{P}((s_{t+1}, m_{t+1})|o, b, a)\\&= \frac{\sum_{s_t} \mathbb{P}(o|(s_t, m_{t+1}), a)\mathbb{P}((s_{t+1}, m_{t+1})|(s_t, m_{t+1}), a)\mathbb{P}((s_t, m_{t+1})|b)}{\sum_{s_{t+1}, m_{t+1}} \sum_{s_t} \mathbb{P}(o|(s_t, m_{t+1}), a)\mathbb{P}((s_{t+1}, m_{t+1})|(s_t, m_{t+1}), a)\mathbb{P}((s_t, m_{t+1})|b)}.\end{aligned}$$

□

Proof of Theorem 1

We start from the end of time horizon $t = T$. For any reachable belief vector $b \in \tilde{B}_T$, we can show that in each iteration of Algorithm 1, the probability of sampling b at time $t = T$ is strictly greater than 0. Suppose b is reachable through the path

$$(b_0, a_0, o_0) \to (b_1, a_1, o_1), ..., \to (b_{T-1}, a_{T-1}, o_{T-1}) \to b_T = b.$$

If we denote f as the smallest non-zero element in F, then in i^{th} iteration of Algorithm 1 for any i,

$$\mathbb{P}(\{b \text{ is sampled in iteration } i\}) \geq (\epsilon f)^T > 0.$$

From the definition of $\hat{V}_T$ in Algorithm 1 we can see,

$$\begin{aligned}&\mathbb{P}(\{V_T(b) - \hat{V}_T^{i+1}(b) > 0\})\\\leq&\mathbb{P}(\{b \text{ is not in } B_T^i\})\\=&\mathbb{P}(\{\text{None of the first } i \text{ iterations has sampled } b\})\\\leq&(1 - (\epsilon f)^T)^i \to 0, \text{ as } i \to \infty.\end{aligned}$$

Thus, $\hat{V}_T^i(b)$ converges to $V_T(b)$ in probability for any $b \in \tilde{B}_T$.

Next, we use induction to show that $\hat{V}_t^i(b)$ converges to $V_t(b)$ in probability for any $b \in \tilde{B}_t$ for all $t \leq T$. In Algorithm 1, it is easy to see that, by applying the backward induction,

$$\hat{V}_t^i(b) = \max_a \sum_{s \in S} b(s)\{r(s, a) + \sum_{o \in O} \sum_{s' \in S} \gamma F(s, a, o)\hat{V}_{t+1}^i(\Lambda(b|a, o))\}, \ \forall b.$$

Then, at time $t < T$, for any belief vector $b \in \tilde{B}_t$, a sufficient condition such that $\hat{V}_t^i(b)$ converges to $V_t(b)$ will be $\hat{V}_{t+1}^i(\Lambda(b|a, o))$ converges to $V_{t+1}(\Lambda(b|a, o))$ for any action a and observation o, i.e., $\hat{V}_{t+1}^i(b')$ converges to $V_{t+1}(b')$ for all b' reachable at time $t + 1$ from b at time t. Starting from V_T, we have already shown $\hat{V}_T^i(b)$ converges to $V_T(b)$ in probability for all reachable belief vector b at time T. Thus, we conclude that $\hat{V}_t^i(b)$ converges to $V_t(b)$ in probability for any $b \in \tilde{B}_t$ for all $t \leq T$.

□

Proof of Proposition 3

Prove by construction. Consider an MPOMDP model $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \lambda)$, where two POMDP models $\mathcal{M}_m = (S, b_0, A, P^m, O, F^m, r^m)$ for $m = 1, 2$ have a same state space, observation space, and action space

$$S = \{s_1, s_2\},\ O = \{o_1, o_2\},\ A = \{a_1, a_2\}$$

but different transition and observation probabilities

$$P^1(a_1) = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix} F^1(a_1) = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix},$$

$$P^1(a_2) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} F^1(a_2) = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix},$$

$$P^2(a_1) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} F^2(a_1) = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix},$$

$$P^2(a_2) = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix} F^2(a_2) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

and different reward functions

$$a_1 : r(s_1, a_1, o_1) = 2, r(s_1, a_1, o_2) = 0, r(s_2, a_1, o_1) = 0, r(s_2, a_1, o_2) = 1$$

$$a_2 : r(s_1, a_2, o_1) = 1, r(s_1, a_2, o_2) = 0, r(s_2, a_2, o_1) = 0, r(s_2, a_2, o_2) = 2.$$

We consider the time horizon to be $t = 0, 1$ and set the model weights to be $\lambda_1 = \lambda_2 = 0.5$. For any belief vector b of $\mathcal{M}$, we write

$$b = (1 - b^1, b^1, 1 - b^2, b^2),$$

where b^1 is the belief in state s_2 in $\mathcal{M}_1$ and b^2 is the belief in state s_2 in $\mathcal{M}_2$. For any α-vector α_t at time $t = 0, 1$, we write

$$\alpha_t = (\alpha_t^1, \alpha_t^2) = (\alpha_t^1(0), \alpha_t^1(1), \alpha_t^2(0), \alpha_t^2(1))$$

where α_t^1 is the α-vector in $\mathcal{M}_1$, $\alpha_t^1(0)$, $\alpha_t^1(1)$ are the values of α_t^1 at $b^1 = 0$ and $b^1 = 1$; and similarly for α_t^2. We can use the exact solution method to find the set of all α-vectors at time $t = 1$:

$$\mathcal{A}_1 = \{(1.6, 0.8, 1.2, 0.6), (0.7, 1.4, 0.9, 1.8)\}.$$

Now, apply Algorithm 1. Suppose in the first iteration we sample two belief vectors at time $t = 0$, which are $b_0^1 = (0.25, 0.25, 0.25, 0.25)$ and $b_0^2 = (0.5, 0, 0, 0.5)$; and then we sample action a_1 and observation o_1, resulting two belief vectors at time $t = 1$, which are $b_1^1 = (0.13, 0.37, 0.29, 0.21)$ and $b_1^2 = (0.07, 0.6, 0.03, 0.3)$.

We then can identify $(0.7, 1.4, 0.9, 1.8) \in \mathcal{A}_1$ as the only non-dominated α-vector at time $t = 1$, and two non-dominated α-vectors

$$(2.019, 2.631, 2.529, 3.021), \ (2.22, 2.28, 1.56, 2.94)$$

at time $t = 0$. In the next iteration, suppose we sample one more belief vector $b_0^3 = (0, 0.5, 0.5, 0)$ at time $t = 0$ and $b_1^3 = (0.45, 0.05, 0.45, 0.05)$ at time $t = 1$ following action a_1 and observation o_1. Then, using three sampled belief vectors b_1^1, b_1^2, b_1^3, we can identify all two α-vectors in $\mathcal{A}_1$ as non-dominated α-vectors at time $t = 1$.

At time $t = 0$, using three sampled belief vectors b_0^1, b_0^2, b_0^3, we can find three non-dominated α-vectors at time $t = 0$, which are

$$(2.019, 2.631, 2.529, 3.021), (2.93, 1.57, 2.19, 2.31), (2.48, 2.32, 2.34, 1.26).$$

In other words, in the first iteration, we have

$$\hat{\mathcal{A}}_0^1 = \{(2.019, 2.631, 2.529, 3.021), (2.22, 2.28, 1.56, 2.94)\}$$

and in the second iteration, we have

$$\hat{\mathcal{A}}_0^2 = \{(2.019, 2.631, 2.529, 3.021), (2.93, 1.57, 2.19, 2.31), (2.48, 2.32, 2.34, 1.26)\}.$$

Now, consider the belief point $b = (0.45, 0.05, 0, 0.5)$ at time $t = 0$:

$$\hat{V}_0^2(b) = 2.552 < 2.583 = \hat{V}_0^1(b),$$

i.e., the lower bound estimate of V_0 after the second iteration is smaller than the estimate after the first iteration at b. $\square$