Submitted to
manuscript

Ambiguous Partially Observable Markov Decision Processes: Structural Results and Applications

Soroush Saghafian

Harvard Kennedy School, Harvard University, Cambridge, MA

Markov Decision Processes (MDPs) and their generalization, Partially Observable MDPs (POMDPs), have been widely studied and used as invaluable tools in dynamic stochastic decision-making. However, two major barriers have limited their application for problems arising in various practical settings: (a) computational challenges for problems with large state or action spaces, and (b) ambiguity in transition probabilities, which are typically hard to quantify. While several solutions for the first challenge, known as "curse of dimensionality," have been proposed, the second challenge remains unsolved and even untouched in the case of POMDPs. We refer to the second challenge as the "curse of ambiguity," and address it by developing a generalization of POMDPs termed Ambiguous POMDPs (APOMDPs). The proposed generalization not only allows the decision maker to take into account imperfect state information, but also tackles the inevitable ambiguity with respect to the correct probabilistic model. Importantly, this paper extends various structural results from POMDPs to APOMDPs. Such structural results can guide the decision maker to make robust decisions when facing model ambiguity. Robustness is achieved by using α-maximin expected utility (α-MEU), which (a) differentiates between ambiguity and ambiguity attitude, (b) avoids the over conservativeness of traditional maximin approaches widely used in robust optimization, and (c) is found to be suitable in laboratory experiments in various choice behaviors including those in portfolio selection. The structural results provided also help to handle the "curse of dimensionality," since they significantly simplify the search for an optimal policy. Furthermore, we provide an analytical performance guarantee for the APOMDP approach by developing a bound for its maximum reward loss due to model ambiguity. To generate further insights into how APOMDPs can help to make better decisions, we also discuss specific applications of APOMDPs including machine replacement, medical decision-making, inventory control, revenue management, optimal search, sequential design of experiments, bandit problems, and dynamic principal-agent models.

Key words: POMDP; unknown probabilities; model ambiguity; structural results; control-limit policies.
History: Version: June 22, 2015

1. Introduction

Markov Decision Processes (MDPs) have been widely used for optimizing Markovian systems in which two main assumptions hold: (1) the state of the system is completely known/observable at each decision epoch, and (2) the state transitions can be probabilistically defined. Partially Observable MDPs (POMDPs) extend MDPs by relaxing the first assumption: POMDPs consider the case where the system's state is not completely observable but there exist observations/signals which yield probabilistic beliefs about the hidden state, if the second assumption above holds. However, the second assumption is unrealistic in most applications, and significantly limits the applicability of POMDPs in realistic problems. In such problems, one might have access to some data, and to develop a POMDP, must first estimate core state and observation transition probabilities. This often comes with estimation errors and leaves the decision maker with inevitable model

misspecification/ambiguity. We refer to this challenge as the *curse of ambiguity,* and address it by relaxing the assumption (1) above. Hence, this paper extends POMPDs to a new dynamic decision-making framework that allows the decision maker to consider both imperfect state information and ambiguity with respect to the correct probabilistic model. We term this new framework as *Ambiguous POMDP (APOMDP).*

To address the curse of ambiguity, we assume that the decision maker simultaneously faces (a) non-probabilistic ambiguity (also known as *Knightian uncertainty*) about the true model, and (b) probabilistic uncertainty or risk given the true model[1]. As Arrow (1951) (p. 418) states: *"There are two types of uncertainty: one as to the hypothesis, which is expressed by saying that the hypothesis is known to belong to a certain class or model, and one as to the future events or observations given the hypothesis, which is expressed by a probability distribution."* Indeed, in our framework, the decision maker is faced with Knightian uncertainty regarding the true model, while under each potential model, he has a certain probabilistic view of how observations and the core system state evolve over time. This provides a distinction between *ambiguity* (lack of knowledge about the true probabilistic model) and *risk* (probabilistic consequences of decisions under a known model).

Another important factor in dealing with ambiguity is the distinction between the ambiguity set of a decision maker (DM, "he" hereafter) and his attitude toward ambiguity. The former refers to characterization of a DM's subjective beliefs (the set of possible probabilistic models) while the latter refers to his taste (the degree of his desire for ambiguity). Given an ambiguity set, the *robust optimization* or *maximin expected utility* (MEU) theory assumes complete aversion to ambiguity and uses the so-called maximin or Wald's criterion by maximizing utility with respect to the worst possible member of the ambiguity set. This, however, typically results in overly conservative decisions (see, e.g., Delage and Mannor (2010) and Xu and Mannor (2012)). Moreover, it is not consistent with several studies that find that the inclusion of *ambiguity seeking* features is behaviorally meaningful. For instance, Bhidé (2000) performs a survey of entrepreneurs and finds that they exhibit a very low level of ambiguity aversion, and Heath and Tversky (1991) demonstrate that individuals who feel competent are in favor of ambiguous situations.

In this paper, to (a) avoid overly conservative outcomes, (b) distinguish between ambiguity and ambiguity attitude, and (c) include more meaningful behavioral aspects, we utilize a generalization of the robust optimization approach and allow the DM to take into account both the worst possible outcome (representing ambiguity aversion) and the best possible outcome (representing ambiguity seeking). The preferences under this criterion are called α-MEU preferences (with "multiple-priors"), and are axiomatized in Ghiradato et al. (2004) (see also Marinacci (2002)). They are found to be suitable for modeling various choice behaviors including those in portfolio

[1] See, e.g., Stoy (2011), for an axiomatic treatment of statistical decision-making under these conditions.

selection (see, e.g., Ahn et al. (2007)).

The fundamental results in allowing for both optimistic and pessimistic views of the world (in a static setting) were communicated by Hurwitcz and Arrow in early 1950s (see, e.g., Arrow and Hurwicz (1997), and Hurwicz (1951a,b)). They postulated four axioms that a choice operator must satisfy, and demonstrated that under complete ignorance, one can restrict attention merely to the extreme outcomes (i.e., the best and the worst). This result constructed a family of utility functions for a DM under ambiguity including linear combinations of the best and worst outcomes, as we consider in this paper.

Since (a) the α-MEU criterion includes Wald's criterion (maximin) as a special case (when the weight assigned to the best possible outcome is zero), and (b) our work allows for incomplete dynamic information, the framework we develop in this paper extends the successful stream of studies on robust MDPs (see, e.g., Nilim and El Ghaoui (2005), Iyengar (2005), Wiesemann et al. (2013)) in two main aspects: (1) it prevents overly conservative decisions by allowing for a controllable "pessimism factor" that can take values in $[0, 1]$, unlike the robust optimization framework where it is constrained to be one (see also the related effort in studies such as Delage and Mannor (2010), Xu and Mannor (2012), and Perakis and Roels (2008) to reduce conservatism). One immediate benefit is related to more realistic behavioral aspects of decision-making discussed earlier. However, perhaps more importantly, our results show that if the DM is hypothetically allowed to optimize his pessimism factor so as to minimize his reward loss when facing model ambiguity, he should choose a mid-range value, i.e., a value that is neither zero nor one. (2) By allowing for incomplete information about the core state (implementing a POMDP rather than a MDP framework), our work is also applicable in several applications where obtaining perfect information regarding the system's state is impossible (e.g., medical decision-making for diseases where tests are subject to false-positive and false-negative errors). To the best of our knowledge, our work is among the very first to allow for both incomplete state information and model ambiguity, both of which are inevitable in many real-world applications[2].

Another challenge in dynamic programming in general, and in MDPs and POMDPs in particular, is the well-known *curse of dimensionality*. It refers to the computational challenges in solving large-scale and challenging dynamic programs. One successful method mainly used for MPDs is to use approximate dynamic programming and other related approximation techniques (see, e.g., Bertsekas and Tsitsiklis (1996), de Farias and Van Roy (2003), Si et al. (2004), Veatch (2013), and the references therein). A separate stream of research that is widely used for MDPs attempts to develop meta-structural results (see, e.g., Smith and McCardle (2002), and the references therein). There are also some limited results in this second vein for POMDPs (see, e.g., Lovejoy (1987b)

[2] We will discuss a variety of such applications in Section 6.

and Rieder (1991)). One of the main contributions of our work is to extend such meta-structural results from POMDPs to APOMDPs.

Specifically, after developing the APOMDP approach and presenting some of its basic properties including contraction mapping of its Bellman operator (on a Banach space) and convergence of a finite-horizon setting to that of an infinite-horizon, we show that unlike the seminal result of Sondik (1971) who proved the convexity (and piecewise-linearity) of the value function for POMDPs (in finite-horizon settings), the APOMDP value function is not always convex: *model ambiguity can cause non-convexity*. Importantly, however, we provide sufficient conditions for the APOMDP value function to be piecewise-linear and convex. Thus, our result builds a bridge between APOMDPs and POMDPs by extending the prominent result of Sondik (1971) from POMDPs to APOMDPs. This, in turn, allows for a similar method of computing the value function as well as the optimal policy in APOMDPs to those developed for POMDPs. Furthermore, using the Blackwell ordering (Blackwell (1951a)) and a variation of the Blackwell-Sherman-Stein sufficiency theorem (Blackwell (1951a, 1953, 1951b), Stein (1951)), we establish the connection of the required condition for the convexity of an APOMDP value function to a notion of *model informativeness* in the "cloud" of models considered by the DM. We also clarify the connection between our result and a different way of handling model misspecification, in which probabilistic beliefs (i.e., information states) are distorted using a martingale process.

We then generate insights into the conditions required to guarantee the convexity of optimal policy regions in the APOMDP framework. The existence of convex policy regions is an important advantage, since it significantly simplifies the search for an optimal policy. We then shed light on the conditions required for an APOMDP value function to be monotone in the belief state space using *Total Positivity of Order 2 (TP_2)* ordering. We do so by showing that monotonicity of an APOMDP value function is indeed preserved under both pessimism and optimism (under some conditions), and hence, under the APOMDP Bellman operator.

We also provide a performance guarantee for the APOMDP approach by bounding the maximum reward loss of a DM who is facing model ambiguity but uses the APOMDP approach compared to an imaginary DM who is fully informed of the true model. Our result allows the DM to adopt an appropriate ambiguity set (i.e., a set of possible models) so as to achieve a required performance guarantee. Through a representative numerical example, we then show that the APOMDP approach is indeed robust to model misspecification. More importantly, we show that the proposed APOMDP approach provides stronger policies than those provided by traditional maximax or maximin criteria: *it is better to choose a mid-level pessimism factor* than the zero or one extreme cases. Using the Hausdorff distance between policy regions obtained by using the best pessimism level and those in a close neighborhood of it, we then provide insights into the robustness of an APOMDP optimal policy to the value of the DM's pessimism level. Doing so, we demonstrate the

equivalence of policy regions under close pessimism levels.

We next discuss a variety of applications of APOMDPs including medical decision-making, inventory control, dynamic pricing and revenue management, optimal search, sequential design of experiments and bandit problems, and dynamic principal-agent models. We argue that while POMDPs are widely used for such applications, the unambiguous knowledge about the core state and observation transition probabilities is an unrealistic assumption in most cases. Since APOMDPs extend POMDPs by relaxing this assumption, they provide a widely useful framework to make more realistic and robust decisions. This is achieved by reducing the reliance on a specific probabilistic model[3].

To also illustrate the advantages of the meta-structural results provided in the paper, we discuss in detail one specific application of APOMDPs: the class of machine replacement problems. The literature on this class of problems assumes a perfect knowledge on deterioration probabilities, while in real-world there exists considerable amount of ambiguity with respect to such probabilities. Thus, we use our proposed APOMDP framework to allow for such ambiguity and shed light on conditions required for the existence of *control-limit* policies. Using a technique for ordering belief points on lines within the underlying simplex, we then provide a novel technique for approximating the control-limit thresholds.

Finally, we briefly discuss a connection between APOMDPs and non-zero-sum dynamic stochastic games with perfect information and an uncountable state space. While several key studies are available for such games (see, e.g., Whitt (1980), Nowak (1985), Nowak and Szajowski (1999), Simon (2007)), various technical challenges remain unsolved, and we leave it to future research to develop further structural results for APOMDPs using a game-theoretical perspective.

The rest of the paper is organized as follows. In Section 2, we briefly review the related studies. The APOMDP framework is presented in Section 3. Section 4 presents the structural properties of APOMDPs, and Section 5 provides a performance guarantee. Section 6 discusses various applications of POMDPs, and Section 7 makes a connection between APOMDPs and stochastic games. Section 8 concludes the paper. All the proofs are presented in Online Appendix A.

2. Literature Review

An important property for dynamic stochastic optimization under incomplete information shown in Aoki (1965), Astrom (1965), and Bertsekas (1976) is that the belief state provides a sufficient statistic of the complete history. This allows formulating the problem using dynamic programming based on the belief state, and establishes a significant connection between MDPs and POMDPs.

[3] We have specifically observed such benefits after implementing and testing the APOMDP framework proposed in this paper in a real-world medical decision-making problem faced by physicians in Mayo Clinic. Further details can be provided upon request.

Drake (1962) is typically known for developing the first POMDP model in the literature. However, the most prominent structural results such as piecewise-linearity and convexity of the value function were first shown by Sondik (1971). Such results were later extended by Smallwood and Sondik (1973) and Sondik (1978) for finite and infinite-horizon problems, respectively. Some other structural results such as monotonicity of the value function in the belief state were later shown by Lovejoy (1987b) and Rieder (1991). Another important structural result is regarding the convexity of policy regions in POMDPs, which is discussed in Lovejoy (1987a). One of our main goals in this paper is to extend such structural results from POMDPs to APOMDPs, and shed light on conditions under which such extensions are possible.

Another stream of research deals with computational methods for solving POMDPs. Since the belief state in a POMDP can take infinitely many values, the POMDP, in its general form, is a very hard problem to solve by the usual methods of value or policy iteration that are typically used for MDPs; for a general POMDP, it is shown that computing an optimal policy is PSPACE-complete (Papadimitriou and Tsitsiklis (1987)) and finding an ϵ-optimal policy is NP-hard (Lusena et al. (2001)). Therefore, some papers discuss methods which discretize the belief space (see, for a reference, Lovejoy (1991b)), remove redundant pieces of the value function over the belief space (see, for a reference, Monahan (1982)), develop computationally feasible bounds (see, e.g., Lovejoy (1991a)), take a geometric approach based on the Minkowski sum of convex polytopes (Zhang (2010)), or focus on specific machine maintenance or other well-structured problems in an attempt to find structural results such as the existence of control-limit policies that simplify the search for optimal policies (see, e.g., Eckles (1968), Ehrenfeld (1976), Grosfeld-Nir (1996, 2007), Monahan (1982), Ross (1971), Wang (1977), White (1977, 1979), Krishnamurthy and Wahlberg (2009), Krishnamurthy (2011)). Finally, many other computer science papers attempt to find numerical solutions to POMDPs. Some examples of these include the "one-pass algorithm" (Sondik (1971)), "linear-support solution" (Cheng (1988)), "witness algorithm" (Cassandra et al. (1994)), "incremental pruning method" (Zhang and Liu (1996)), "region-based incremental pruning" (Feng and Zilberstein (2004)), and "bounded finite state controller" (Poupart and Boutilier (2004)).

Another line of research that is more related to our work (mainly developed in the economics theory literature) deals with decision-making under ambiguity or model misspecification. A stream of research by Hansen and Sargent discusses model ambiguity and illuminates ways for creating robust frameworks (see, e.g., Hansen and Sargent (2007, 2008, 2012)). The α-MEU preferences that we use in this paper is discussed and axiomatized in Marinacci (2002) and Ghiradato et al. (2004), and found to be suitable in laboratory experiments for modeling choice behaviors in applications such as portfolio selection (see, e.g., Ahn et al. (2007)). The α-MEU criterion generalizes the MEU preferences in which the DM only considers the worst-case outcome. MEU preferences are widely used in robust optimization and specifically in robust MDPs (see, e.g., Nilim and El Ghaoui (2005),

Iyengar (2005), Wiesemann et al. (2013)), but typically result in overly conservative policies (see, e.g., Delage and Mannor (2010) and Xu and Mannor (2012)). The α-MEU criterion avoids this conservatism by considering both the best and the worst outcomes. Furthermore, the α-MEU criterion allows for a differentiation between the DM's ambiguity and ambiguity attitude. This differentiation is also achieved in smooth model of decision-making under ambiguity proposed by Klibanoff et al. (2005) and Klibanoff et al. (2009), where smoothness is obtained by considering a "second order" belief that reflects the DM's subjective belief about the potential models. However, this requires consideration of all ambiguous outcomes and comes with extra computational burden, especially if used for POMDPs (which are already computationally complex).

To the best of our knowledge, this paper is among the very first to develop a POMDP-type framework under ambiguity. Considering (a) the wide-range of applications of POMDPs in various fields including medicine, biology, operations research, economics, computer science, and engineering, among others, and (b) the fact that in most applications, model parameters cannot be exactly estimated (due to factors such as insufficient data, disagreement among experts, etc.), we believe the APOMDP framework and related structural results developed in this paper are of high value for many applications. A similar effort can be found in Itoh and Nakamura (2007) and Hansen and Sargent (2007). However, both of these papers differ from our work in two main ways: (a) they do not develop meta-structural results (e.g., convexity, monotonicity, etc.) that can simplify the search for optimal policies (as is of our goals in this paper), and (b) the decision-making criterion and the framework developed in them is significantly different from our proposed APOMDP model.

In closing this section, we also note that some papers assume perfect state information, but pursue the use of data and partial distributional information in a dynamic way to overcome model ambiguity (e.g., through learning). For this stream of research, we refer interested readers to Saghafian and Tomlin (2015) (and the references therein), in which data and partial distributional information are dynamically used (via entropy maximization) to reduce the DM's ambiguity over time. In this paper, similar to the literature on robust MDPs (see, e.g., Nilim and El Ghaoui (2005), Iyengar (2005), Wiesemann et al. (2013)), the goal is not to reduce or overcome ambiguity (e.g., through learning). Instead, we focus on dynamic decision-making when ambiguity is inevitable. Unlike the literature on robust MDPs, however, we (a) allow for unobservable states, and (b) consider both the best and worst outcomes to avoid over-conservatism, and thereby achieve policies that are behaviorally more relevant.

3. The APOMDP Framework

A discrete-time, infinite-horizon, discounted reward APOMDP with finite actions and states is an extension of the classical POMDP, and can be defined by the tuple (α, β, $\mathscr{S}$, $\mathscr{O}$, $\mathscr{A}$, $\mathscr{G}$, $\mathscr{P}$, $\mathscr{R}$). In this definition (1) α and β denote the pessimism level and the discount rate, respectively. (2)

$\mathscr{S}=\{1,2,\cdots,n\}$, $\mathscr{O}=\{1,2,\cdots,k\}$, and $\mathscr{A}=\{1,2,\cdots,l\}$ are finite sets representing state space, observation space, and action space, respectively. (3) $\mathscr{G}=\{g^a\in\mathbb{R}^n:\forall a\in\mathscr{A}\}$ is the set of immediate rewards, where g^a is a vector with ith element being the immediate reward of being at state $i\in\mathscr{S}$ when action $a\in\mathscr{A}$ is taken. (4) $\mathscr{P}$ and $\mathscr{R}$ are the *ambiguity sets* with respect to core state and observation transition probabilities, respectively[4].

To construct a single ambiguity set and simplify our notation, we consider $\mathscr{P}\times\mathscr{R}$, assume it is a finite set, and denote by $m\in\mathscr{M}\triangleq\{1,2,\ldots,|\mathscr{P}\times\mathscr{R}|\}$ an index that uniquely represents its members[5]. In this view, we consider $\mathscr{M}$ as a "cloud" of models (a new ambiguity set), with m being a specific model in the "cloud." Thus, associated with each model m is a set of the form $P_m\times R_m$ with P_m and R_m denoting the set of state and observation transition probabilities under model m, respectively. In this setting, $P_m=\{P_m^a:a\in\mathscr{A}\}$, where for each $a\in\mathscr{A}$ $P_m^a=[p_{ij}^a(m)]_{i,j\in\mathscr{S}}$ is an $n\times n$ matrix with $p_{ij}^a(m)=Pr\{j|i,a,m\}$ denoting the probability that the system's core state moves to j from i under action a and model m. Similarly, $R_m=\{R_m^a:a\in\mathscr{A}\}$, where for each $a\in\mathscr{A}$, $R_m^a=[r_{jo}^a(m)]_{j\in\mathscr{S},o\in\mathscr{O}}$ is an $n\times k$ matrix with $r_{jo}^a(m)=Pr\{o|j,a,m\}$ denoting the probability of observing o under action a and model m when the core state is j.

For any real-valued finite set Ξ, we let Π_Ξ denote the probability simplex induced by Ξ. In particular, we let $\Pi_\mathscr{S}$ denote the $(n-1)$-simplex representing the probability belief space about the system's state. We denote by $\mathscr{T}\triangleq\{0,1,\ldots,T\}$ the decision epochs, where T is the time horizon. We also let $\mathscr{I}\triangleq[0,1]$, and assume $\alpha\in\mathscr{I}$ and $\beta\in\mathscr{I}\setminus\{1\}$.

If $\mathscr{M}$ was a singleton with its only member being m (i.e., under a complete confidence about the model), the optimal reward and policy for any $t\in\mathscr{T}$ and $\pi\in\Pi_\mathscr{S}$ could be obtained by a traditional POMDP Bellman equation (along with the terminal condition $V_0^m(\pi)=\pi' g_0$ for some $g_0\in\mathbb{R}^n$):

$$V_t^m(\pi)=\max_{a\in\mathscr{A}}\Big\{\pi' g^a+\beta\sum_{o\in\mathscr{O}}Pr\{o|\pi,a,m\}V_{t-1}^m(T(\pi,a,o,m))\Big\},\tag{1}$$

where all the vectors are assumed to be in column format, " $'$ " represents a transpose, g^a represents a vector of size n with elements being expected single-period reward of being at each state under action a, $Pr\{o|\pi,a,m\}=\sum_i\sum_j\pi_i p_{ij}^a(m)r_{jo}^a(m)$ is the probability of observing o under belief π, action a, and model m. The belief updating operator $T:\Pi_\mathscr{S}\times\mathscr{A}\times\mathscr{O}\times\mathscr{M}\to\Pi_\mathscr{S}$ in (1) is defined by the Bayes' rule (in the matrix form):

$$T(\pi,a,o,m)=\frac{\big(\pi' P_m^a R_m^a(o)\big)'}{Pr\{o|\pi,a,m\}},\tag{2}$$

[4] It should be noted that we focus on ambiguity with respect to core state and observation transition probabilities. This is because in robust dynamic programming settings under model ambiguity and expected discounted reward, the reward function can be assumed to be certain without loss of generality (see, e.g., Iyengar (2005)).

[5] The assumption that $\mathscr{P}\times\mathscr{R}$, and hence $\mathscr{M}$, is finite is only made for the ease of indexing, and is not a restrictive assumption; the majority of the results in this paper can be easily extended to cases with an infinite or even uncountable set $\mathscr{M}$. It should be also noted that any continuous set of transition probabilities can be approximated via finite sets with any required precision. Thus, one can always consider a finite set $\mathscr{M}$ as a close approximation to a continuous one.

where $R^a_m(o) \triangleq \text{diag}(r^a_{1o}(m), r^a_{2o}(m), \ldots, r^a_{no}(m))$ is the diagonal matrix made of the oth column of R^a_m. Letting

$$\hat{h}_{t-1}(\pi, a, m) \triangleq \sum_{o \in \mathscr{O}} Pr\{o|\pi, a, m\} V^m_{t-1}(T(\pi, a, o, m)) \tag{3}$$

denote the "reward-to-go" function, the POMDP optimality equation for model m can be written as

$$V^m_t(\pi) = \max_{a \in \mathscr{A}} \Big\{ \pi' g^a + \beta \hat{h}_{t-1}(\pi, a, m) \Big\}. \tag{4}$$

Using the preliminaries above, we now consider the APOMDP case. In particular, we note that in an APOMDP, the DM is faced with model misspecification and only ambiguously (not even probabilistically) knows m: he only knows that $m \in \mathscr{M}$. Hence, the reward and policy cannot be simply obtained from (4). To obtain the optimality equation of APOMDP, we consider the α-MEU criterion as follows. We let $\alpha \in \mathscr{I}$ denote the pessimism factor, and denote by $\underline{m}_{t-1}$ and $\overline{m}_{t-1}$ the worst and best-case models (values of $m \in \mathscr{M}$) with respect to the $(t-1)$-periods expected reward-to-go function, respectively:

$$\underline{m}_{t-1}(\pi, a, \alpha) \triangleq \arg\min_{m \in \mathscr{M}} \; h_{t-1}(\pi, a, m, \alpha), \tag{5}$$

$$\overline{m}_{t-1}(\pi, a, \alpha) \triangleq \arg\max_{m \in \mathscr{M}} \; h_{t-1}(\pi, a, m, \alpha). \tag{6}$$

In this setting,

$$h_{t-1}(\pi, a, m, \alpha) \triangleq \sum_{o \in \mathscr{O}} Pr\{o|\pi, a, m\} V_{t-1}(T(\pi, a, o, m), \alpha), \tag{7}$$

and $V_{t-1}(\pi, \alpha)$ denotes the decision maker's reward with $t-1$ periods to go[6]. Using this notation, the DM's reward and policy under the APOMDP framework can be obtained by solving the following equation (along with the terminal condition $V_0(\pi, \alpha) = \pi' g_0$ for some $g_0 \in \mathbb{R}^n$):

$$V_t(\pi, \alpha) = \max_{a \in \mathscr{A}} \Big\{ \pi' g^a + \beta \big[\alpha\, h_{t-1}(\pi, a, \underline{m}_{t-1}(\pi, a, \alpha), \alpha) + (1-\alpha)\, h_{t-1}(\pi, a, \overline{m}_{t-1}(\pi, a, \alpha), \alpha) \big] \Big\}. \tag{8}$$

We refer to (8) as the finite-horizon APOMDP Bellman equation, and call the policy obtained by solving it α-Hurwicz[7] or H^α for short[8]. For notational convenience, we define the utility function

$$U_t(\pi, \alpha, a) = \pi' g^a + \beta \big[\alpha\, h_{t-1}(\pi, a, \underline{m}_{t-1}(\pi, a, \alpha), \alpha) + (1-\alpha)\, h_{t-1}(\pi, a, \overline{m}_{t-1}(\pi, a, \alpha), \alpha) \big], \tag{9}$$

which allows us to concisely write the finite-horizon APOMDP Bellman equation as

$$V_t(\pi, \alpha) = \max_{a \in \mathscr{A}} U_t(\pi, \alpha, a). \tag{10}$$

[6] Note that since the best and worst-case models are chosen independently of the previous periods best and worst-case model selections, our setting also satisfies the rectangularity assumption discussed by Epstein and Schneider (2003) in their recursive multiple-prior setting, and used by Nilim and El Ghaoui (2005) and Iyengar (2005) for robust MDPs.

[7] We adopt this terminology to emphasize the seminal work of Hurwicz (1951a) in decision-making under complete ignorance.

[8] It should be clear that H^0 and H^1 policies are the widely used maximax and maximin (Wald's) criteria, respectively.

However, when more convenient, we write the finite-horizon APOMDP Bellman equation (8) in an operator form. To this end, we let $\mathscr{B}^\alpha$ denote the set of real-valued bounded functions defined on $\Pi_{\mathscr{S}} \times \{\alpha\}$, and define the operator $\mathfrak{L}^\alpha : \mathscr{B}^\alpha \to \mathscr{B}^\alpha$ based on (7)-(8) such that $V_t = \mathfrak{L}^\alpha V_{t-1}$ for $t = 1, 2, \cdots, T$. The following lemma shows that the operator $\mathfrak{L}^\alpha$ is a contraction mapping with modulus β on the Banach space $(\mathscr{B}^\alpha, d^\alpha)$, where for any $V, W \in \mathscr{B}^\alpha$, the metric d^α is defined as $d^\alpha(V,W) \triangleq \sup_{\pi \in \Pi_{\mathscr{S}}} |V(\pi,\alpha) - W(\pi,\alpha)|$. This will enable us to establish a fixed-point result for APOMDPs.

LEMMA 1 **(Contraction Mapping Bellman Operator)**. *For all $\alpha \in \mathscr{I}$, the APOMDP Bellman operator $\mathfrak{L}^\alpha$ is a contraction mapping with modulus β on the Banach space $(\mathscr{B}^\alpha, d^\alpha)$. That is, for any $V, W \in \mathscr{B}^\alpha$: $d^\alpha(\mathfrak{L}^\alpha W, \mathfrak{L}^\alpha V) \leq \beta\, d^\alpha(W,V)$.*

The following result uses Lemma 1, and sheds light on the connection between finite-horizon and infinite-horizon APOMDPs by using the *Banach's Fixed-Point Theorem*. To consider infinite-horizon APOMDPs, we let $T = \infty$ and denote the infinite-horizon APOMDP value function by $V_\infty(\pi,\alpha)$.

PROPOSITION 1 **(APOMDP Convergence)**. *For all $\pi \in \Pi_{\mathscr{S}}$ and $\alpha \in \mathscr{I}$, $V_\infty(\pi,\alpha)$ is the unique solution to $\mathfrak{L}^\alpha V_\infty(\pi,\alpha) = V_\infty(\pi,\alpha)$. Furthermore, for all $\pi \in \Pi_{\mathscr{S}}$ and $\alpha \in \mathscr{I}$, $\lim_{t\to\infty} V_t(\pi,\alpha) = V_\infty(\pi,\alpha)$, where the convergence is uniform (in d^α).*

REMARK 1 (**Stochastic Games with Perfect Information**). It is noteworthy that the APOMDP framework introduced above (in both finite and infinite-horizon cases) can also be viewed as a non-zero-sum sequential stochastic game with perfect information and an uncountable state space. We briefly discuss this connection in Section 7.

REMARK 2 (**Dynamic Consistency**). As some studies including Ghirardato et al. (2008) discuss, the presence of ambiguity in dynamic settings might lead to violations of dynamic consistency in preferences. It should be noted that this issue is not of first order importance in our work, because motivated by various applications (see, e.g., Section 6) our goal is to consider a DM who is facing both ambiguity and imperfect state information, and prescribe a policy which is (a) *behaviorally meaningful*, and (b) *effective* in dealing with ambiguity. The policy that is obtained by solving the APOMDP program introduced above achieves both of these goals. In particular, while in earlier sections we discussed the literature addressing the behavioral aspects of considering both the best and worst-case outcomes, in later sections we show (both analytically and numerically) the effectiveness of an APOMDP policy in dealing with ambiguity. However, as we discuss in more detail in Online Appendix B, dynamic consistency in our framework, if needed, can be obtained in at least two ways. First, attention can be restricted to specific values of pessimism level (e.g., $\alpha = 0$, $\alpha = 1$, and some ranges including them) for which dynamic consistency of preferences is preserved.

Second, the DM can be allowed to dynamically adjust his pessimism level. While we only consider a static pessimism factor in this paper, our results in Section 5 show that small changes in the pessimism level does not affect the adopted policy by the DM.

4. Basic Structural Results for APOMDPs

4.1. Convexity

Solving the APOMDP functional equation (8) can be complex in general. In particular, in contrast to the seminal result of Sondik (1971) who proved the convexity of the value function for POMDPs, we observe that the APOMDP value function is not always convex in $\pi \in \Pi_{\mathscr{S}}$. Hence, unlike POMDPs, the APOMDP value function does not always admit the desirable form $V_t(\pi,\alpha) = \max_{\psi\in\Psi_{t,\alpha}}\{\pi'\psi\}$ for some finite set of vectors $\Psi_{t,\alpha}$. We illustrate this through the following example.

EXAMPLE 1 (**Non-Convex Value Function**). Consider an APOMDP with $n=k=l=3$ (i.e., three states, three observations, and three actions). Suppose there are three models ($m=1$, $m=2$, and $m=3$) with transition probabilities given for each model in a separate row below.

$$P_1^1=\begin{pmatrix}0.3&0.5&0.2\\0.1&0.6&0.3\\0.2&0.3&0.5\end{pmatrix}\ P_1^2=\begin{pmatrix}0.1&0.6&0.3\\0.5&0.2&0.3\\0.4&0.3&0.3\end{pmatrix}\ P_1^3=\begin{pmatrix}0.3&0.3&0.4\\0.1&0.7&0.2\\0.5&0.3&0.2\end{pmatrix}\ R_1^1=\begin{pmatrix}0.4&0.3&0.3\\0.1&0.8&0.1\\0.1&0.2&0.7\end{pmatrix}\ R_1^2=\begin{pmatrix}0.1&0.3&0.6\\0.4&0.3&0.3\\0.2&0.1&0.7\end{pmatrix}\ R_1^3=\begin{pmatrix}0.5&0.2&0.3\\0.4&0.4&0.2\\0.3&0.1&0.6\end{pmatrix}$$

$$P_2^1=\begin{pmatrix}0.3&0.6&0.1\\0.3&0.6&0.1\\0.1&0.1&0.8\end{pmatrix}\ P_2^2=\begin{pmatrix}0.2&0.7&0.1\\0.5&0.2&0.3\\0.3&0.2&0.5\end{pmatrix}\ P_2^3=\begin{pmatrix}0.2&0.5&0.3\\0.1&0.5&0.4\\0.2&0.2&0.6\end{pmatrix}\ R_2^1=\begin{pmatrix}0.5&0.3&0.2\\0.4&0.3&0.3\\0.1&0.1&0.8\end{pmatrix}\ R_2^2=\begin{pmatrix}0.1&0.4&0.5\\0.5&0.2&0.3\\0.3&0.1&0.6\end{pmatrix}\ R_2^3=\begin{pmatrix}0.2&0.2&0.6\\0.4&0.1&0.5\\0.6&0.2&0.2\end{pmatrix}$$

$$P_3^1=\begin{pmatrix}0.2&0.6&0.2\\0.2&0.4&0.4\\0.2&0.4&0.4\end{pmatrix}\ P_3^2=\begin{pmatrix}0.6&0.1&0.3\\0.4&0.4&0.2\\0.5&0.1&0.4\end{pmatrix}\ P_3^3=\begin{pmatrix}0.1&0.8&0.1\\0.2&0.7&0.1\\0.2&0.2&0.6\end{pmatrix}\ R_3^1=\begin{pmatrix}0.3&0.3&0.4\\0.4&0.3&0.3\\0.1&0.1&0.8\end{pmatrix}\ R_3^2=\begin{pmatrix}0.2&0.4&0.4\\0.6&0.2&0.2\\0.8&0.1&0.1\end{pmatrix}\ R_3^3=\begin{pmatrix}0.2&0.2&0.6\\0.5&0.2&0.3\\0.6&0.2&0.2\end{pmatrix}$$

Also, let $g_0=(1.0,1.1,1.0)'$, $g^1=(1.5,1.8,1.6)'$, $g^2=(1.7,1.4,1.5)'$, $g^3=(1.6,1.7,1.5)'$, $T=3$, and $\beta=0.9$.

Figure 1 illustrates the value function with $\alpha=0.95$ at $t=3$ for various belief points $\pi=(\pi_1,\pi_2,\pi_3=1-\pi_1-\pi_2)\in\Pi_{\mathscr{S}}$. As Figure 1 shows, the value function is not convex: *model ambiguity causes non-convexity*. However, in what follows we show that, under a condition on the ambiguity set defined below, the seminal result of Sondik (1971) for POMDPs can be extended to APOMDPs.

DEFINITION 1 (**Dominant Worst-Case Member**). The ambiguity set $\mathscr{M}$ is said to have a dominant worst-case member if $\underline{m}_t(\pi,a,\alpha)$ is constant in π.

PROPOSITION 2 **(Piecewise-Linearity and Convexity)**. *If the ambiguity set $\mathscr{M}$ has a dominant worst-case member, then $V_t(\pi,\alpha)$ for any finite t is piecewise-linear and convex in π, and hence admits $V_t(\pi,\alpha)=\max_{\psi\in\Psi_{t,\alpha}}\{\pi'\psi\}$ for some finite set of vectors $\Psi_{t,\alpha}$.*

Proposition 2 extends the seminal result of Sondik (1971) from POMDPs to APOMDPs. We note that (a) the condition in Proposition 2 is only on the worst-case scenario: no condition on the best-case is required, and (b) Definition 1 allows the dominant worst-case member to change

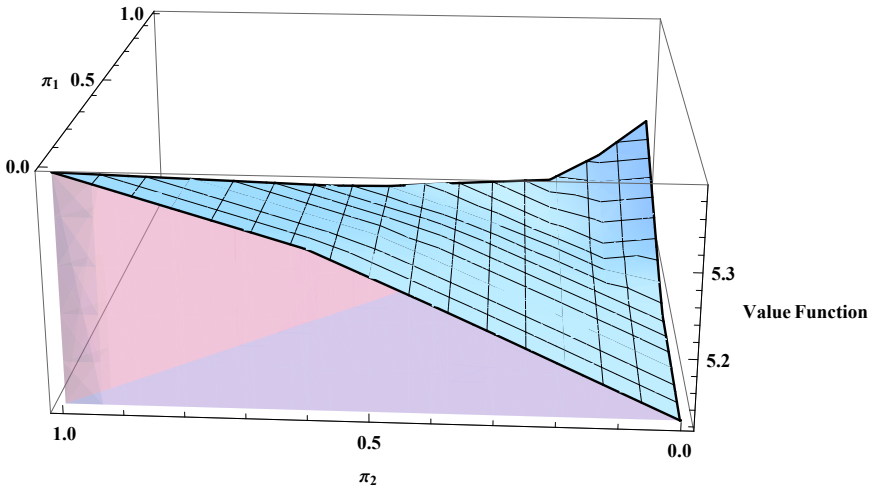


Figure 1 The APOMDP value function of Example 1.

dynamically over time, or based on the action or conservatism level but not the DM's belief. That is, to yield a convex value function (of π), Proposition 2 requires the adversary (but not necessarily the ally) part of the nature to act independently of the DM's belief, π. For instance, the DM may share its belief only with the ally part of the nature. Importantly, however, the ambiguity set can have a dominant worst-case member under various other situations (even when the belief is shared with the adversary part of the nature), and hence the condition in Proposition 2 is not that restrictive. An important example is related to the notion of *model informativeness* introduced below.

Definition 2 (**Model Informativeness**). A model $m^* \in \mathscr{M}$ is said to be less informative than another model $m \in \mathscr{M}$ under action $a \in \mathscr{A}$, if there exists a k by k transition probability kernel Q_m^a such that $P_{m^*}^a R_{m^*}^a = P_m^a R_m^a Q_m^a$.

The above definition of model informativeness is equivalent to Blackwell Ordering (Blackwell (1951a, 1953)), which is often referred to as *information garbling* in the economics of information literature (see, e.g., Marschak and Miyasawa (1968)). The property above can be understood by noting that $P_m^a R_m^a$ is a matrix of signals or conditional probabilities of the form $[Pr\{i|o\}]_{i\in\mathscr{S},o\in\mathscr{O}}$ under model m. The above definition describes that model $m^* \in \mathscr{M}$ is less informative than model m, if m^* provides signals that are weaker than those under m, in that the signals (about core states) received under m^* are only "garbled" (through channel/transformation Q_m^a) versions of signals received under m. That is, one could retrieve the signals under m^* if s/he had access to signals under m. Thinking of this signals as outputs of statistical experiments under m^* and m, we first state the following variation of the Blackwell-Sherman-Stein sufficiency theorem[9] to connect model informativeness in our framework with convex stochastic ordering[10] (denoted by $\preceq_{cx}$) of the

[9] The result is originally due to Blackwell (1951a, 1953, 1951b), Stein (1951).

[10] For more details, see, e.g., Chapter 3. of Shaked and Shanthikumar (2007).

posterior likelihood distributions defined by the operator in (2). This will then allow us to connect model informativeness to the existence of a dominant worst-case model.

LEMMA 2 **(Model Informativeness Ordering)**. *Suppose model $m_1 \in \mathscr{M}$ is less informative than model $m_2 \in \mathscr{M}$ under an action $a \in \mathscr{A}$. If observation O is considered as a random variable, then:*

$$T(\pi, a, O, m_1) \preceq_{cx} T(\pi, a, O, m_2). \tag{11}$$

The above lemma is equivalent to the following statement: model $m_1 \in \mathscr{M}$ being less informative than model $m_2 \in \mathscr{M}$ under an action $a \in \mathscr{A}$ results in

$$\mathbb{E}_{O|\pi,a,m_1}[f(T(\pi, a, O, m_1)] \leq \mathbb{E}_{O|\pi,a,m_2}[f(T(\pi, a, O, m_2)], \tag{12}$$

for any real-valued convex function f defined on $\Pi_{\mathscr{S}}$. In other words, any utility maximizer with a convex utility f, that depends on the posterior belief, prefers the statistical experiment governed by m_2 than one governed by m_1 (under action a). Using this result, we can state the following:

PROPOSITION 3 **(Model Informativeness and Dominant Worst-Case Member)**. *Fix $\alpha \in \mathscr{I}$ and suppose that under each action $a \in \mathscr{A}$, one of the models denoted by $m^*(a)$ is less informative than all the other models in $\mathscr{M}$. Then $\mathscr{M}$ has a dominant worst-case member. Furthermore, $\underline{m}_t(\pi, a, \alpha) = m^*(a)$ for all $t \in \mathscr{T}$.*

An important aspect of the above result is that the required condition only depends on the DM's ambiguity set and not his ambiguity attitude. This is because the proposed APOMDP framework allows for a separation between ambiguity and ambiguity attitude as discussed in Section 1. Moreover, when one of the models is less informative than the rest, Proposition 3 shows that the adversary part of the nature acts independently of the DM's belief (even if the DM shares his belief), and hence, the ambiguity set will have a dominant worst-case member. The following remark describes yet another important aspect of Proposition 3.

REMARK 3 (**Martingale Distorted Beliefs**). It should be noted that convex stochastic ordering is closely related to martingale representations (see, e.g., Theorem 3.A.4 of Shaked and Shanthikumar (2007)), and so (11) and Proposition 3 can be viewed as follows. Suppose at each period, the DM uses model $m^*(a)$ as his approximate/reference model under action $a \in \mathscr{A}$, but uses a martingale distortion of his future belief about the hidden state (conditioned on his current belief), since he does not fully trust his approximate model $m^*(a)$. If the "cloud" of models, $\mathscr{M}$, is built indirectly in this way (as opposed to directly considering different state and observation transition kernels), then our results above indicate that $\mathscr{M}$ will still have a worst-case member, and most of our main results will still hold. The idea of using martingale distortions (without commitment to previous period distortions) to represent model misspecification has appeared in Hansen and Sargent (2007, 2008).

We now turn our attention to the properties of the optimal APOMDP policy. Let the mapping $a_t^* : \Pi_{\mathscr{S}} \times \mathscr{I} \to \mathscr{A}$ denote the optimal APOMDP policy with t periods to go, and define the sets $\Pi^*_{t,a}(\alpha) \triangleq \{\pi \in \Pi_{\mathscr{S}} : a_t^*(\pi, \alpha) = a\}$, which we refer to as *policy regions*.

First, we note that even in a traditional POMDP, the policy regions may not be convex unless some conditions hold (see, e.g., Ross (1971), White (1978), Lovejoy (1987a)). We illustrate through the following example that the same observation holds for APOMDPs.

Example 2 (**Policy Regions**). Consider the APOMDP of Example 1. Figure 2 illustrates the policy regions at $t = 3$ for various levels of pessimism, α. As can be seen from parts (a), (c), and (d) of this figure, the policy regions are not always convex. Moreover, a maximin (robust optimizer) DM (Figure 2 part (d)) will use action 1 unless he is somehow confident that the system is at state 1. Such a DM will use action 1 more than any other optimizer. In contrast, a maximax DM (Figure 2 part (a)) uses action 3 more than any other optimizer. An H^α optimizer (with a mid range α), however, will make a careful balance between using actions 1 and 3.

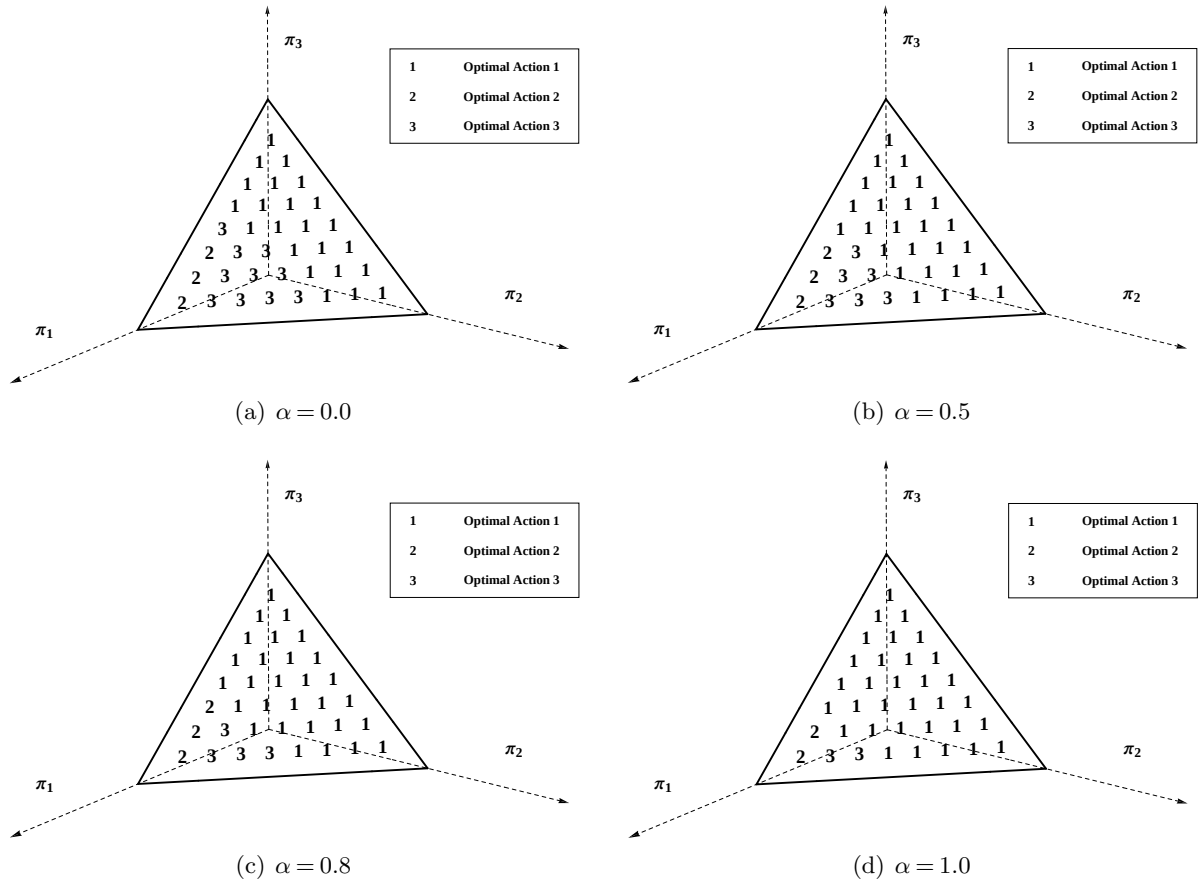
While the policy regions under H^α are not necessarily convex, the following result presents sufficient conditions for their convexity.

Proposition 4 **(Convex Policy Regions)**. *If (a) the ambiguity set $\mathscr{M}$ has a dominant worst-case member, and (b) under an action $a \in \mathscr{A}$ the core state becomes perfectly known under both $\underline{m}_t$ and $\overline{m}_t$, then $\Pi^*_{t,a}(\alpha)$ is a convex set ($\forall \alpha \in \mathscr{I}$).*

Condition (a) of the above proposition holds if, for instance, one of the models is less informative than the rest (Proposition 3). It should be also noted that condition (b) of the above proposition appears in many applications. For instance, in machine replacement problems, machine/production deterioration is typically ambiguous and hard to define through one probabilistic model, making the proposed APOMDP framework an attractive decision-making tool. However, when the machine is replaced or fully inspected, the system's core state becomes observable under any legitimate model (see, e.g., Ross (1971) and White (1978)). In Section 6.1, we will discuss this class of problems in more depth.

4.2. Monotonicity

In this section, we explore conditions under which one can guarantee the monotonicity of the APOMDP value function. Such results allow a DM to gain insights into the structure of the optimal policy without any computational effort. For instance, in Section 6.1, we will use the general monotonicity results developed here to establish the existence of optimal *control-limit* policies for machine replacement problems under ambiguity. We will also discuss a method to effectively approximate the control-limit threshold itself by introducing a technique for ordering beliefs on lines.



(a) $\alpha = 0.0$ (b) $\alpha = 0.5$

(c) $\alpha = 0.8$ (d) $\alpha = 1.0$

Figure 2 Policy regions of the APOMDP for various levels of pessimism, α.

We start by stating the monotonicity of the APOMDP value function in the DM's pessimism level.

PROPOSITION 5 **(Monotonicity: Pessimism Level)**. *The value function $V_t(\pi, \alpha)$ is non-increasing in α ($\forall t \in \mathscr{T}$, $\forall \pi \in \Pi_{\mathscr{S}}$).*

A more important monotonicity result is related to the DM's information state (belief vector). To compare two elements of the information state space $\Pi_{\mathscr{S}}$, one needs to use a stochastic ordering which is preserved under the Bayesian operator (2). Total Positivity of Order 2 (TP_2) is the natural choice for this purpose.

DEFINITION 3 (**Total Positivity of Order 2 (TP_2), Karlin and Rinott (1980)**). Denote by f and g two real-valued ν-variate functions defined on $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2 \times \ldots \times \mathscr{X}_\nu$ where each $\mathscr{X}_i$ is totally ordered. f is said to be larger than or equal to g in (multivariate) Total Positivity of Order 2 sense ($g \preceq_{TP_2} f$) if for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}$: $f(\mathbf{x} \vee \mathbf{y})\, g(\mathbf{x} \wedge \mathbf{y}) \geq f(\mathbf{x})\, g(\mathbf{y})$, where $\vee$ and $\wedge$ are the usual (componentwise max and min) lattice operators. Similarly, for two probability vectors (or multi-dimensional mass functions) $\pi = (\pi_i : i \in \mathscr{S}) \in \Pi_{\mathscr{S}}$ and $\hat{\pi} = (\hat{\pi}_i : i \in \mathscr{S}) \in \Pi_{\mathscr{S}}$, we use the notation $\pi \preceq_{TP_2} \hat{\pi}$, if $\pi_i \hat{\pi}_{\hat{i}} \geq \pi_{\hat{i}} \hat{\pi}_i$ whenever $i \leq \hat{i}$ and $i, \hat{i} \in \mathscr{S}$.

The TP_2 ordering defined above reduces to the Monotone Likelihood Ratio (MLR) ordering

for univariate functions (i.e., when $\nu = 1$), and so is also known as strong MLR ordering (Whitt (1982)). However, it should be noted that, unlike MLR ordering, TP_2 is not reflexive, which causes additional challenges in partially observable systems.

DEFINITION 4 (**Reflexive TP_2 Functions**). A function f is said to be reflexive TP_2 (or, for simplicity, TP_2) if $f \preceq_{TP_2} f$.

DEFINITION 5 (**TP_2 Transition Kernels**). For a given model $m \in \mathscr{M}$, the set of state transition probability kernels $P_m = \{P_m^a : a \in \mathscr{A}\}$ is said to be TP_2, if the function $p_{ij}^a(m) = Pr\{j|i,a,m\}$ defined on $\mathscr{X} = \mathscr{S} \times \mathscr{S}$ is TP_2 for all $a \in \mathscr{A}$ [11]. Similarly, for a given model $m \in \mathscr{M}$, the set of observation transition probability kernels $R_m = \{R_m^a : a \in \mathscr{A}\}$ is said to be TP_2 if the function $r_{jo}^a(m) = Pr\{o|j,a,m\}$ defined on $\mathscr{X} = \mathscr{S} \times \mathscr{O}$ is TP_2 for all $a \in \mathscr{A}$.

We also need to define the set of real-valued TP_2-nondecreasing functions induced by $\Pi_{\mathscr{S}}$.

DEFINITION 6 (**Real-Valued TP_2-Nondecreasing Functions**). The set of real-valued TP_2-nondecreasing functions induced by $\Pi_{\mathscr{S}}$, denoted by $\mathscr{F}_{\Pi_{\mathscr{S}}}$, is the set of all real-valued functions defined on $\Pi_{\mathscr{S}} \times \mathscr{X}_1 \times \mathscr{X}_2 \times \ldots \times \mathscr{X}_\nu$ (for some arbitrary $\nu \in \mathbb{N}$ and sets $\mathscr{X}_1, \ldots, \mathscr{X}_\nu$) such that $f \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ if $f(\pi, \ldots, .) \leq f(\hat{\pi}, \ldots, .)$, whenever $\pi \preceq_{TP_2} \hat{\pi}$ and $\pi, \hat{\pi} \in \Pi_{\mathscr{S}}$.

In a POMDP, it is known that $V_t^m(\pi) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ under some conditions (see, e.g., Proposition 1 of Lovejoy (1987b), or Theorem 2.4 of Rieder (1991)). We now extend this result to APOMDPs by showing that $V_t(\pi, \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ under some similar conditions. We start with the following lemma, which shows that the set $\mathscr{F}_{\Pi_{\mathscr{S}}}$ is closed under both pessimism and optimism.

LEMMA 3 (**Closedness of $\mathscr{F}_{\Pi_{\mathscr{S}}}$ under Pessimism and Optimism**). *If $h_t(\pi, a, m, \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ for all $m \in \mathscr{M}$, then $h_t(\pi, a, \underline{m}_t(\pi, a, \alpha), \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ and $h_t(\pi, a, \overline{m}_t(\pi, a, \alpha), \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$.*

We also establish the closedness of $\mathscr{F}_{\Pi_{\mathscr{S}}}$ under observation-based expectation operators.

LEMMA 4 (**Closedness of $\mathscr{F}_{\Pi_{\mathscr{S}}}$ under Expectation**). *Let $g(o|\pi) \in \Pi_{\mathscr{O}}$ be a probability mass function such that $g(o|\pi_1) \preceq_{TP_2} g(o|\pi_2)$ whenever $\pi_1 \preceq_{TP_2} \pi_2$ $(\pi_1, \pi_2 \in \Pi_{\mathscr{S}})$. If (i) $f(\pi, o) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$, and (ii) $f(\pi, o)$ is nondecreasing in $o \in \mathscr{O}$, then $\mathbb{E}_{g(o|\pi)}[f(\pi, o)] \in \mathscr{F}_{\Pi_{\mathscr{S}}}$.*

Using the above lemmas, we first show that under some conditions the set $\mathscr{F}_{\Pi_{\mathscr{S}}}$ is closed under APOMDP value iteration. This, in turn, will alow us to establish important monotonicity results for the APOMDP value function. In what follows, we let $\uparrow\mathbb{R}^n$ denote the set of all vectors in $\mathbb{R}^n$ with an ascending order of elements[12].

PROPOSITION 6 (**Monotonicity Preservation in APOMDP**). *Suppose the set of kernels P_m and R_m are TP_2 for all $m \in \mathscr{M}$.*

[11] This is equivalent to all the second-order minors of matrix $P_m^a = [p_{ij}^a(m)]_{i,j \in \mathscr{S}}$ being non-negative for all $a \in \mathscr{A}$.

[12] For instance, $(1, 2, \ldots, n-1, n)' \in {}^{\uparrow}\mathbb{R}^n$, $(1, 1, \ldots, 1, 1)' \in {}^{\uparrow}\mathbb{R}^n$, but $(1, 2, \ldots, n, n-1)' \notin {}^{\uparrow}\mathbb{R}^n$

(i) If $V_{t-1}(\pi,\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$*, then* $h_{t-1}(\pi,a,m,\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$ *for all* $m\in\mathscr{M}$*.*
(ii) If $V_{t-1}(\pi,\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$ *and* $g^a\in{}^{\uparrow}\mathbb{R}^n$ *for all* $a\in\mathscr{A}$*, then* $V_t(\pi,\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$*.*

Finally, the following theorem presents conditions for both finite and infinite-horizon settings under which the value function of an APOMDP is monotonic. It provides a generalization for Proposition 1 of Lovejoy (1987b) and Theorem 4.2 of Rieder (1991) which establish similar results for traditional POMDPs. We again highlight that the structural results we have established in this and previous section have important implications in a variety of applications, some of which we will discuss in Section 6.

THEOREM 1 **(Monotonicity in APOMDP)**. *Suppose the set of kernels* P_m *and* R_m *are* TP_2 *for all* $m\in\mathscr{M}$ *and* $g^a\in{}^{\uparrow}\mathbb{R}^n$ *for all* $a\in\mathscr{A}$*.*
(i) If $T<\infty$ *and* $g_0\in{}^{\uparrow}\mathbb{R}^n$*, then* $V_t(\pi,\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$ *for all* $t\in\mathscr{T}$ *and* $\alpha\in\mathscr{I}$*.*
(ii) If $T=\infty$*, then* $V_\infty(\pi,\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$ *for all* $\alpha\in\mathscr{I}$*.*

5. Performance Guarantee and Robustness of the APOMDP Policy

We now first explore the effectiveness of the policy obtained by solving an APOMDP in dealing with ambiguity. In particular, we consider the optimal APOMDP policy (H^α), and derive a bound for the maximum reward loss that may occur when there is model ambiguity and H^α is implemented compared to when the correct model is completely known and an optimal POMDP policy is used (i.e., the absolute best-case under no model ambiguity). In this way, we provide a *performance guarantee* for using H^α when facing model ambiguity. As we will see, this will also enable the DM to investigate whether the ambiguity set he is using is "tight" enough. We then explore the robustness of the H^α policy to variations in the pessimism level, α.

To provide a performance guarantee, we need some preliminary definitions and results. First, we need a measure for the "tightness" of the ambiguity set.

DEFINITION 7 (**ϵ-Tightness**). The ambiguity set $\mathscr{M}$ is said to be ϵ-tight if for any two $m_1,m_2\in\mathscr{M}$:

$$|p^a_{ij}(m_1)r^a_{jo}(m_1)-p^a_{ij}(m_2)r^a_{jo}(m_2)|\le\epsilon\qquad\forall i,j\in\mathscr{S},\forall o\in\mathscr{O},\forall a\in\mathscr{A}. \tag{13}$$

An APOMDP is said to be ϵ-tight if its ambiguity set is ϵ-tight[13].

Using the notion of ϵ-tightness defined above, we now bound the maximum difference in the vector of conditional observation probabilities $Pr(o|\pi,a,m)\triangleq(Pr\{o|\pi,a,m_1\}:o\in\mathscr{O})$ caused by model ambiguity.

LEMMA 5 (**$\mathcal{L}_1$-Norm Bound**). *For any ϵ-tight APOMDP:*

$$||Pr(o|\pi,a,m_1)-Pr(o|\pi,a,m_2)||_1\le\xi\qquad\forall m_1,m_2\in\mathscr{M},\forall\pi\in\Pi_{\mathscr{S}},\forall a\in\mathscr{A},$$

[13] By this definition, a 0-tight APOMDP is a POMDP. It should be also noted that a larger "cloud" of models (i.e., a larger $\mathscr{M}$) typically (but not necessarily) results in a weaker level of tightness.

where $||\cdot||_1$ *is the* $\mathcal{L}_1$*-norm,* $\xi = \min\{k\,n\,\epsilon, 2\}$*,* $n = |\mathscr{S}|$*, and* $k = |\mathscr{O}|$*.*

When the DM is facing ambiguity, his belief about the core state (i.e., his information state $\pi \in \Pi_{\mathscr{S}}$) at any decision epoch might be distorted compared to when he knows the exact model. We next present a similar result to that of Lemma 5, but by considering the case where the DM's belief state is distorted.

LEMMA 6 **(Belief Distortion)**. *For any two belief states* $\pi, \hat{\pi} \in \Pi_{\mathscr{S}}$*:*

$$||Pr(o|\pi, a, m) - Pr(o|\hat{\pi}, a, m)||_1 \leq ||\pi - \hat{\pi}||_1 \qquad \forall m \in \mathscr{M}, \forall a \in \mathscr{A}.$$

In addition, if the APOMDP is ϵ*-tight, then:*

$$||Pr(o|\pi, a, m_1) - Pr(o|\hat{\pi}, a, m_2)||_1 \leq \xi + ||\pi - \hat{\pi}||_1 \qquad \forall m_1, m_2 \in \mathscr{M}, \forall a \in \mathscr{A},$$

where ξ *is defined in Lemma 5.*

We next consider the effect of belief distortion by assuming that a model $m \in \mathscr{M}$ is indeed the true model. For generality, we allow the DM to follow any arbitrary policy (within the class of deterministic and Markovian policies) $\eta : \Pi_{\mathscr{S}} \times \mathscr{I} \to \mathscr{A}$, and denote by $V_t^{m,\eta}(\pi)$ the reward obtained under such policy when m is the true model and the belief is $\pi \in \Pi_{\mathscr{S}}$. Assuming model m is the true model, we let $V_t^{m,\eta}(\hat{\pi})$ denote the reward obtained under the same actions used to calculate $V_t^{m,\eta}(\pi)$, but when the belief is $\hat{\pi} \in \Pi_{\mathscr{S}}$ instead of $\pi \in \Pi_{\mathscr{S}}$ (i.e., a distorted belief).

Moreover, without loss of generality and for clarity, we assume $g_0 = 0$ in the rest of this section[14].

LEMMA 7 **(Max Reward Loss - Distorted Belief)**. *Under any policy* η*:*

$$|V_t^{m,\eta}(\pi) - V_t^{m,\eta}(\hat{\pi})| \leq \frac{(1-\beta^{t+1})\,\overline{g}}{(1-\beta)^2}\,||\pi - \hat{\pi}||_1 \qquad \forall \pi, \hat{\pi} \in \Pi_{\mathscr{S}}, \forall m \in \mathscr{M},$$

where $\overline{g} = \max_{a \in \mathscr{A}} ||g^a||_\infty$ *(with* $||\cdot||_\infty$ *denoting the* $\mathcal{L}_\infty$*-norm).*

Next, we need to bound the maximum difference between the reward obtained from the APOMDP versus that obtained from a POMDP, when the DM uses the same policy in both: when using a fixed policy, what is the maximum reward loss caused by model ambiguity? To provide the answer, we let $V_t^{\eta}(\pi, \alpha)$ denote the APOMDP value function under policy η, and compare it with the corresponding POMDP value function under η and model m, denoted by $V_t^{m,\eta}(\pi)$.

LEMMA 8 **(Max Reward Loss - Arbitrary Policy)**. *If the APOMDP is* ϵ*-tight, then under any policy* η*:*

$$|V_t^{m,\eta}(\pi) - V_t^{\eta}(\pi, \alpha)| \leq \frac{\bar{\xi}\,\beta\,(3-\beta)(1-\beta^t)\,\overline{g}}{(1-\beta)^3} \qquad \forall m \in \mathscr{M}, \forall \pi \in \Pi_{\mathscr{S}}, \forall t \in \mathscr{T}, \forall \alpha \in \mathscr{I},$$

where $\bar{\xi} = \min\{k\,n\,\epsilon, \frac{3\,(1-\beta)}{3-\beta}\}$*,* $n = |\mathscr{S}|$*, and* $k = |\mathscr{O}|$*.*

[14] Extending the results to $g_0 \neq 0$ is straightforward and is left to reader.

Finally, we present our main performance guarantee result by bounding the maximum reward loss that may occur by following the H^α policy instead of the optimal policy of the no-ambiguity case. This bounds the maximum reward loss of the H^α policy when evaluated in (and compared to) any of the POMDP models in the ambiguity set.

THEOREM 2 **(Max Reward Loss - Optimal Policy)**. *If the APOMDP is ϵ-tight, then*

$$V_t^m(\pi) - V_t^{m,H^\alpha}(\pi) \leq \frac{3\,\bar{\xi}\,\beta\,(3-\beta)(1-\beta^t)\,\overline{g}}{(1-\beta)^3} \qquad \forall m \in \mathscr{M}, \forall \pi \in \Pi_{\mathscr{S}}, \forall t \in \mathscr{T}, \forall \alpha \in \mathscr{I},$$

where $\bar{\xi}$ is defined in Lemma 8.

The bound provided in Theorem 2 is tight. For instance, it goes to zero as $\epsilon \to 0$. Theorem 2 also allows a DM (who is facing model ambiguity but follows H^α) to determine if his ambiguity set is tight enough: for a desired performance guarantee (maximum reward loss), he can determine the required "tightness" of the ambiguity set (regardless of its cardinality). This insight is established in the following result, where it is assumed $\overline{g} \neq 0$ and $\beta \neq 0$ to avoid trivial cases.

COROLLARY 1 **(Performance Guarantee)**. *Suppose, facing model ambiguity, the DM follows the H^α policy over t periods. If $\mathscr{M}$ is chosen so that it is ϵ-tight for some $\epsilon \leq \overline{\epsilon}$, where*

$$\overline{\epsilon} = \frac{(1-\beta)^3\,\delta_t}{3\,k\,n\,\beta\,(3-\beta)(1-\beta^t)\,\overline{g}} \qquad (\overline{g} \neq 0, \beta \neq 0),$$

then a performance guarantee (maximum reward loss) of δ_t is ensured.

The above results provide a performance guarantee for following the H^α policy when facing model ambiguity. We generate more insights into the robustness of such policy under model ambiguity through the following experiment.

EXAMPLE 3 (**Robustness under** H^α). We use the APOMDP of Example 1, and to gain insights into the performance of the H^α policy, we consider different DMs and uniformly distribute them over the simplex $\Pi_{\mathscr{S}}$. The location of each DM represents his starting belief point. We do this by creating grids of 0.05 in the belief simplex and by locating a DM on each grid point. This results in considering the performance of $\binom{20+3-1}{3-1} = 231$ different DMs[15]. Every time, we give a specific value to α and ask all the DMs to follow the H^α policy with the given value of α. For each DM in this setting, we calculate the average reward loss by assuming that any of the models in the ambiguity set can be the true model. Since the DMs do not have any knowledge about which model is the true model, we assume each of the models can be the true model with an equal chance. We then consider the total average reward loss (due to model ambiguity) among all the DMs, when they follow the H^α policy (obtained by solving an APOMDP for each DM), as our performance metric. An important point to notice is that, for each DM, we are able to calculate the H^α policy as well

[15] The number of distinct nonnegative integer solutions satisfying $\sum_{i=1}^n x_i = c$ is $\binom{c+n-1}{n-1}$.

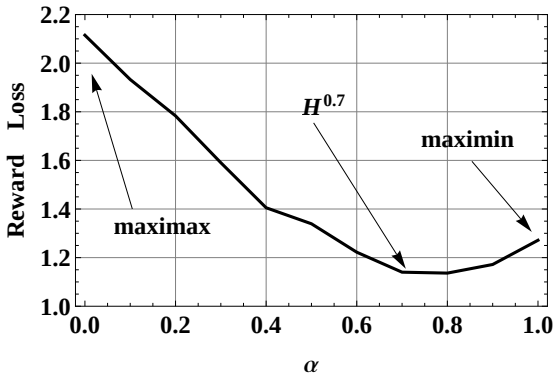


Figure 3 The reward loss with model ambiguity under the H^α policy for various levels of α. Dynamic versions of widely used maximin ($\alpha=1$) and maximax ($\alpha=0$) policies are dominated by the H^α policy for some mid-level α (e.g., $\alpha\in[0.6,0.9]$).

as its average reward loss exactly (no approximation). Figure 3 illustrates the performance for various values of α. From our experiment, we gain the following insights. (a) Maximin ($\alpha=1$) and maximax ($\alpha=0$) policies are dominated by H^α policies with some mid-level α (e.g., $\alpha\in[0.6,0.9]$). Hence, the H^α policy is a valuable generalization of policies such as maximin and maximax as it provides more robustness. (b) The maximin criterion widely used in robust optimization performs better than the maximax one, but as discussed in (a), both can be improved by using a mid-level pessimism factor. (c) While there exists an optimal level α^* (=0.7 in this example), the performance of H^α is quite robust when α is in a range close to α^*. Hence, we observe that, even if a DM's pessimism level is not exactly α^* but is close to it, his policy performs well[16].

The latter observation (part (c)) can be established more formally. In fact, we can show that if a DM's pessimism level is not exactly α^* but is "close" to it, then his policy regions are no different than those defined by α^*. To this end, we use the following way of measuring the distance between two sets, which is essentially the maximum distance of a set to the nearest point in the other set.

Definition 8 (**Hausdorff Distance**). Consider two non-empty sets $\Xi_1, \Xi_2 \subset \mathbb{R}^n$. The Hausdorff distance between Ξ_1 and Ξ_2 (with $\mathcal{L}_\infty$-norm) is

$$d_H(\Xi_1,\Xi_2) \triangleq \max\Big\{\sup_{\xi_1\in\Xi_1}\inf_{\xi_2\in\Xi_2}||\xi_1-\xi_2||_\infty, \sup_{\xi_2\in\Xi_2}\inf_{\xi_1\in\Xi_1}||\xi_1-\xi_2||_\infty\Big\}. \tag{14}$$

It should be noted that $d_H(\Xi_1,\Xi_2)=0$ if, and only if, Ξ_1 and Ξ_2 have the same closures. In particular, if Ξ_1 and Ξ_2 are closed sets, then $d_H(\Xi_1,\Xi_2)=0$ if, and only if, $\Xi_1=\Xi_2$.

Using the above definition, we can now show the following result, which establishes the equivalence between optimal policy regions under α^* and those of any α in a neighborhood of α^*.

Proposition 7 **(Robustness in Pessimism Level)**. *For all $t\in\mathscr{T}$ and α^* in the interior of $\mathscr{I}$, there exists $\epsilon>0$ such that*

[16] While these insights are presented here for a specific example (setting of Example 1), we have observed similar insights from tests under various other different settings. So the result seems to hold widely.

$$\max_{a\in\mathscr{A}:\,\Pi^*_{t,a}(\alpha^*)\neq\emptyset} d_H\big(\Pi^*_{t,a}(\alpha^*),\Pi^*_{t,a}(\alpha)\big)=0, \tag{15}$$

and $\Pi^*_{t,a}(\alpha)\neq\emptyset$ *whenever* $\Pi^*_{t,a}(\alpha^*)\neq\emptyset$*, for all* $\alpha\in\mathscr{I}$ *satisfying* $|\alpha^*-\alpha|<\epsilon$*.*

6. Applications of APOMDPs

In this section, we consider a few important applications of APOMDPs. We start by introducing the class of machine replacement problems with ambiguous deteriorations. We next show how the structural results provided in the earlier sections can be used for this class of problems. Specifically, we (a) prove the existence of *control-limit policies* for this class of problems (under some conditions), and (b) provide an effective method for approximating the control-limits, which significantly reduces the computational difficulty in characterizing the H^α policy. We then briefly discuss various other applications of APOMDPs.

6.1. Machine Replacement Problems with Ambiguous Deterioration

A variety of papers in the literature use POMDPs to study machine replacement problems (see, e.g., White (1977, 1978, 1979), Wang (1977), Ross (1971), Maillart (2006), Jin et al. (2011)). An important assumption in such studies is that the deterioration probabilities are completely known. However, this is a strong assumption and is often not encountered in practice. The proposed APOMDP approach provides a natural framework to relax such an assumption, and provide robust policies that do not heavily rely on a given probability transition matrix. This is an important advantage considering that deterioration probabilities are hard (and often impossible) to quantify in practice.

To gain deeper insights, and show the use of the structural properties provided earlier in this paper, we consider machine replacement problems in which $\mathscr{A}=\{1,2\}$. Considering this special class allows us to gain deeper insights into effective methods for characterizing optimal policies in APOMDPs. We note that, with general number of actions, a simple characterization of the optimal policy may not be achievable even for traditional POMDPs which ignore the underlying ambiguity. For instance, Ross (1971) shows that even for a two-state POMDP, the optimal policy may be complex, and may not have a control-limit structure. Here, we consider the more general class of APOMDPs by allowing for model ambiguity and general number of states, but focus on binary action cases. We present conditions under which a control-limit policy is optimal for any $\alpha\in\mathscr{I}$ (including special cases of robust optimization or maximax approaches introduced earlier). Furthermore, we present a tractable procedure to directly approximate the control-limit thresholds, which significantly reduces the computational difficulty in characterizing the H^α policy.

We start by introducing the class of Binary Action Monotone Machine Replacement (BAMMR) APOMDPs.

DEFINITION 9 (**BAMMR APOMDPs**). An APOMDP is called a Binary Action Monotone Machine Replacement (BAMMR), if it satisfies the following conditions: (a) $\mathscr{A}=\{1,2\}$, (b) the set of kernels P_m^2 and R_m^2 are TP_2 for all $m\in\mathscr{M}$, (c) $p_{ij}^1(m)=\mathbb{1}_{\{j=s(m)\}}$ for all $i,j\in\mathscr{S}$, all $m\in\mathscr{M}$, and some (potentially model-dependent) $s(m)\in\mathscr{S}$, and (d) $g^1,g^2,g^2-g^1\in{}^{\uparrow}\mathbb{R}^n$.

To present structural results for BAMMR APOMDPs, we need the following definition whose power is in yielding monotone optimal policies (see Topkis (1998)).

DEFINITION 10 (TP_2**-Supermodularity**). When $\mathscr{A}=\{1,2\}$, the DM's APOMDP utility defined in (9) is said to be TP_2-supermodular, if $U_t(\pi,\alpha,2)-U_t(\pi,\alpha,1)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$ $(\forall t\in\mathscr{T})$.

LEMMA 9 **(Supermodularity of BAMMR APOMDPs)**. *For any BAMMR APOMDP:*
(i) If $T<\infty$ and $g_0\in{}^{\uparrow}\mathbb{R}^n$, then the DM's utility is TP_2-supermodular.
(ii) If $T=\infty$, then the DM's utility is TP_2-supermodular.

The following result presents sufficient conditions for the existence of optimal control-limit policies for BAMMR APOMDPs. It provides an important extension for the available results on machine replacement problems without model ambiguity (see, e.g., Theorem 4.4. of Rieder (1991), Ross (1971), White (1977, 1978, 1979)).

THEOREM 3 **(Control-Limit Policy)**. *For any BAMMR APOMDP:*
(i) If $T<\infty$ and $g_0\in{}^{\uparrow}\mathbb{R}^n$, then $a_t^*(\pi,\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$ for all $t\in\mathscr{T}$ and $\alpha\in\mathscr{I}$. Furthermore, if $\mathscr{M}$ has a dominant worst-case member, then $\Pi_{t,1}^*(\alpha)$ is a convex set $(\forall t\in\mathscr{T},\ \forall\alpha\in\mathscr{I})$.
(ii) If $T=\infty$, then $a_\infty^*(\pi,\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$ for all $\alpha\in\mathscr{I}$ for all $\alpha\in\mathscr{I}$. Furthermore, if $\mathscr{M}$ has a dominant worst-case member, then $\Pi_{\infty,1}^*(\alpha)$ is a convex set $(\forall\alpha\in\mathscr{I})$.

The above result provides conditions under which a BAMMR APOMDP has an optimal control-limit policy. However, the TP_2 ordering in the above result is stronger than what is needed, and does not help to characterize the threshold surface, since it only induces a partial ordering[17] on $\Pi_{\mathscr{S}}$. However, we can restrict our attention to TP_2 ordering on lines, which will resolve the issue[18]. To this end, let $e_i\in\Pi_{\mathscr{S}}$ represent a vector with a one as the i-th element and zeros elsewhere, denote the convex hull of $e_1,e_2,\cdots,e_{n-1}$ by $\mathcal{C}$, and let $\mathscr{L}(\tilde{\pi})$ be the line in $\Pi_{\mathscr{S}}$ that connects $\tilde{\pi}\in\mathcal{C}$ to e_n:

$$\mathscr{L}(\tilde{\pi})\triangleq\{\pi\in\Pi_{\mathscr{S}}:\ \pi=\lambda\tilde{\pi}+(1-\lambda)e_n,\tilde{\pi}\in\mathcal{C},\lambda\in\mathscr{I}\}.$$

DEFINITION 11 (TP_2 **Ordering on Lines**). Vector $\pi\in\Pi_{\mathscr{S}}$ is said to be less than or equal to vector $\hat{\pi}\in\Pi_{\mathscr{S}}$ in the TP_2 ordering sense on lines (denoted by $\pi\preceq_{TP_2-L}\hat{\pi}$) if $\pi\preceq_{TP_2}\hat{\pi}$ and $\pi,\hat{\pi}\in$

[17] One cannot always compare two members of $\Pi_{\mathscr{S}}$, using the TP_2 ordering, and hence, $[\Pi_{\mathscr{S}},\preceq_{TP_2-L}]$ is a poset.

[18] See Krishnamurthy and Djonin (2009) for similar results in a POMDP application on radar resource management.

$\mathscr{L}(\tilde{\pi})$ (i.e., if π and $\hat{\pi}$ are on the same line connecting e_n to a point in $\tilde{\pi} \in \mathcal{C}$). Moreover, a real-valued function f is said to be TP_2 nondecreasing on lines denoted by $f \in \mathscr{F}^L_{\Pi_{\mathscr{S}}}$, if the condition $\pi \preceq_{TP_2} \hat{\pi}$ in Definition 6 is replaced with $\pi \preceq_{TP_2-L} \hat{\pi}$.

Theorem 3 enables us to state that for any BAMMR APOMDP (a) the optimal policy $a^*_t(\pi,\alpha) \in \mathscr{F}^L_{\Pi_{\mathscr{S}}}$, and (b) for each α, there exists a threshold surface $\Upsilon_t(\alpha)$ that partitions the information space $\Pi_{\mathscr{S}}$ into two individually connected sets[19] such that $a^*_t(\pi,\alpha)=1$ if π is in the first region and $a^*_t(\pi,\alpha)=2$ otherwise[20], even if the ambiguity set $\mathscr{M}$ does not have a dominant worst-case member.

Proposition 8 **(Connectedness)**. *For any BAMMR APOMDP:*
(i) If $T<\infty$ and $g_0 \in {}^{\uparrow}\mathbb{R}^n$, then the policy regions $\Pi^*_{t,1}(\alpha)$ and $\Pi^*_{t,2}(\alpha)$ are both connected sets ($\forall t \in \mathscr{T}$, $\forall \alpha \in \mathscr{I}$).
(ii) If $T=\infty$, then the policy regions $\Pi^*_{\infty,1}(\alpha)$ and $\Pi^*_{\infty,2}(\alpha)$ are both connected sets ($\forall \alpha \in \mathscr{I}$).
(iii) If $\mathscr{M}$ has a dominant worst-case member, then $\Pi^*_{t,1}(\alpha)$ is a convex set ($\forall t \in \mathscr{T}$, $\forall \alpha \in \mathscr{I}$). Thus, the threshold surface $\Upsilon_t(\alpha)$ is convex and almost everywhere differentiable ($\forall t \in \mathscr{T}$, $\forall \alpha \in \mathscr{I}$).

When $a^*_t(\pi,\alpha) \in \mathscr{F}^L_{\Pi_{\mathscr{S}}}$ and the policy regions are connected sets, to characterize the policy regions for a BAMMR APOMDP, an algorithmic procedure can move from e_n to a member of $\mathcal{C}$ (a decreasing direction in the TP_2-L sense) and find the point after which the optimal action changes from 2 to 1 (the control-limit point). If this procedure is repeated for all the members of $\mathcal{C}$, the set of all control-limit points form the threshold surface $\Upsilon_t(\alpha)$. While theoretically sound, this can be computationally intractable. Thus, we next present a technique to effectively approximate $\Upsilon_t(\alpha)$. This will provide an easy-to-calculate method to characterize the policy regions. For simplicity, we present the method for the infinite-horizon case, i.e., to approximate $\Upsilon_\infty(\alpha)$. For the ease of notation, we use $\Upsilon(\alpha)$ instead of $\Upsilon_\infty(\alpha)$ in what follows.

6.1.1. Approximating the Threshold in BAMMR APOMDPs We now present a method to calculate the best linear approximation for the threshold $\Upsilon(\alpha)$ in any BAMMR APOMDP. To this end, consider the vector $\hat{\Upsilon}(\alpha)=(\hat{\Upsilon}_i(\alpha): i \in \mathscr{S}) \in \mathbb{R}^n_+$, and let the corresponding policy defined by it be:

$$a^{\hat{\Upsilon}(\alpha)}(\pi,\alpha)=\begin{cases} 1 : \pi' \hat{\Upsilon}(\alpha) \le 1, \\ 2 : \pi' \hat{\Upsilon}(\alpha) > 1. \end{cases} \tag{16}$$

The choice of 1 in the RHS of (16) and the condition $\hat{\Upsilon}(\alpha) \in \mathbb{R}^n_+$ are both made for uniqueness purposes. In fact, the choice of 1 avoids non-uniqueness that may occur due to scaling, and is without loss of generality. The condition $\hat{\Upsilon}(\alpha) \in \mathbb{R}^n_+$ is added, because in the Euclidean space one can always add a vector with same elements to a vector to make it positive.

[19] A set is connected if it cannot be divided into two disjoint non-empty closed sets.

[20] The infinite-horizon case is similar after setting $t=\infty$.

Now, consider the optimization program

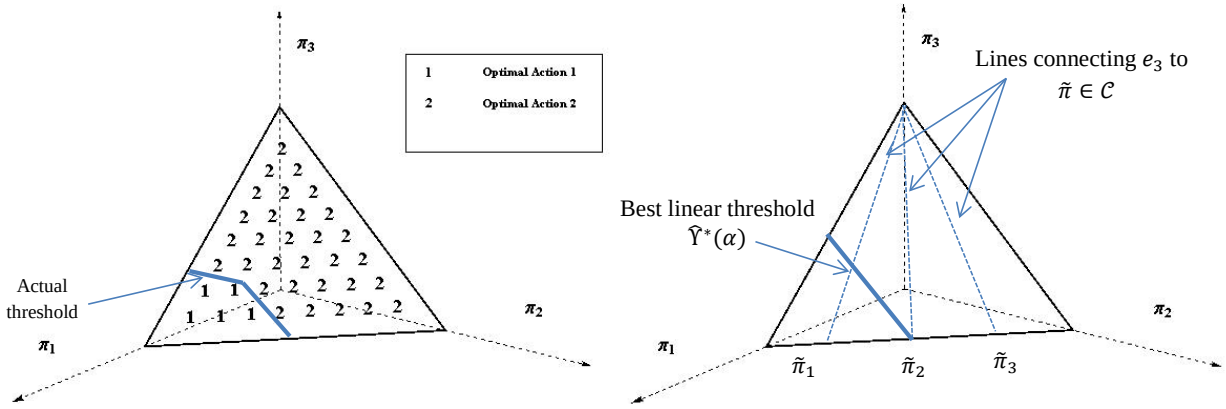
$$\hat{\Upsilon}^*(\alpha) = \arg\max_{\hat{\Upsilon}(\alpha)\in\mathbb{R}^n_+} V_\infty^{\hat{\Upsilon}(\alpha)}(\pi,\alpha)$$
$$\text{s.t.} \quad ||\hat{\Upsilon}(\alpha)||_\infty = \hat{\Upsilon}_n(\alpha), \tag{17}$$

where $V_\infty^{\hat{\Upsilon}(\alpha)}(\pi,\alpha)$ denotes the infinite-horizon BAMMR APOMDP value function under the policy defined by (16). We claim that the above optimization program yields the best linear threshold surface for the BAMMR APOMDP resulting in a TP_2-L nondecreasing policy. To show this claim, we demonstrate that the condition of optimization program (17), which is a "maximum-last-element" requirement, is both necessary and sufficient for characterizing control-limit policies that are $TP2-L$ in any BAMMR APOMDP. Hence, it guarantees (a) inclusion of all the TP_2-L nondecreasing policies, and (b) exclusion of all policies that are not TP_2-L nondecreasing.

PROPOSITION 9 (**TP_2-L Threshold**). *$a^{\hat{\Upsilon}(\alpha)}(\pi,\alpha)\in\mathscr{F}^L_{\Pi_\mathscr{S}}$ if, and only if, $||\hat{\Upsilon}(\alpha)||_\infty=\hat{\Upsilon}_n(\alpha)$.*

REMARK 4 (**Computing the Approximate Threshold**). Based on the result above, program (17) yields the best linear approximation for the BAMMR APOMDP threshold. Solving this program is, however, computationally challenging because it involves computing the APOMDP objective function for various policies. But similar to Krishnamurthy and Djonin (2009) who study the application of a POMDP (not an APOMDP) on radar management, one can use simulation optimization to efficiently solve program (17), thereby characterizing an effective policy for the underlying BAMMR APOMDP. To this end, the objective function of program (9) can be replaced with the expected value of its sample path counterpart (obtained from simulating the APOMDP), creating a stochastic optimization program. Moreover, it should be noted that the constrained program (17) can be easily transferred to an unconstrained one. For instance, one can use the change of variable $\hat{\Upsilon}_n(\alpha)=\big(\tilde{\Upsilon}_n(\alpha)\big)^2$ and $\hat{\Upsilon}_i(\alpha)=\big(\tilde{\Upsilon}_n(\alpha)\sin(\tilde{\Upsilon}_i(\alpha))\big)^2$. Then, the nonnegativity and the maximum-last-element requirements of program (17) are automatically satisfied after the program is written in terms of the new vector $\tilde{\Upsilon}(\alpha)=(\tilde{\Upsilon}_i(\alpha):i\in\mathscr{S})$. Finally, a gradient-based algorithm can be used to efficiently solve the underlying stochastic optimization problem. We refer interested readers to Section III.C of Krishnamurthy and Djonin (2009) for more details about these steps and efficiency of this approach.

EXAMPLE 4 (**Approximating the Threshold**). We consider a BAMMR APOMDP with $n=3$ states and $|\mathscr{M}|=3$ models. We let $g_0=(1.0,1.0,1.1)'$, $g^1=(1.60,1.80,1.89)'$, $g^2=(1.60,1.80,1.90)$ so that $g_0,g^1,g^2,g^2-g^1\in^\uparrow\mathbb{R}^n$. We also let $p^1_{ij}(m)=\mathbb{1}_{\{j=s(m)\}}$ for all $i,j\in\mathscr{S}$, where $s(1)=1$, $s(2)=s(3)=2$. We set $\mathscr{A}=\{1,2\}$, $\alpha=0.4$, and choose the rest of parameters similar to those used in Example 1. The optimal APOMDP policy regions for this setting are shown in Figure 4 (a). The best linear threshold obtained is depicted in Figure 4 (b). As can be seen, the best linear threshold closely approximates the optimal threshold shown in Figure 4 (a). Indeed, the



(a) Actual APOMDP threshold and policy regions (b) Threshold approximation

Figure 4 Threshold approximation for a BAMMR APOMDP.

approximation provides an effective control policy for the BAMMR APODMP problem. Figure 4 (a) also shows that the policy region for action 1, $\Pi^*_{t,1}(\alpha)$, is a convex set, and that the threshold is continuous and almost everywhere differentiable (as predicted by Proposition 8 part (iii)).

6.2. Other Applications of APOMDPs

We now briefly discuss various other applications of APOMDPs.

Medical Decision-Making: Imperfect Tests and Health Transitions. POMDPs are widely used in medical decision-making and related fields. Examples include drug therapy (Lovejoy (1996)), cancer screening (see, e.g., Ayer et al. (2012), Maillart et al. (2008), and Chhatwal et al. (2010)), and biopsy referral decisions (Zhang et al. (2012)). For other examples of using POMDPs in healthcare, we refer to Peek (1999), and the references therein. In most of such applications, the patient's "health state" is not observable mainly because medical tests are imperfect. Thus, one is inclined to use a POMDP approach. However, to do so, the core state and observation transition probabilities need to be estimated from data sets or through methods such as simulation. Typically, such approaches result in estimation errors, leaving the model developer in an ambiguous situation regarding the correctness of the underlying probabilistic model. In fact, in many medical decision-making applications, some actions are not often used in practice, and hence, there is only a very limited data (if any) regarding patient health transition probabilities under such actions. The APOMDP approach developed in this paper allows the decision maker to also take into account the inevitable model misspecifications, and take medical actions that are robust considering the inevitable model ambiguity. This is especially important in healthcare applications, since ignoring model ambiguity can result in wrong medical actions and may have dire consequences on a patient's health.

Stochastic Inventory Control: Record Inaccuracy and Ambiguity in Demand. MDPs are widely used in inventory theory. Recently, a variety of papers have addressed record inaccuracy

in the retail industry, and have developed POMDPs to study inventory control for such systems (see, e.g., DeHoratius et al. (2008)). Typically, in such applications, the demand distribution is not completely known. However, inventory control models generally assume that the demand distribution is completely known, which is a strong assumption and may lead to ineffective control policies (see also Saghafian and Tomlin (2015), and the references therein for more discussions in this vein). The APOMDP approach proposed in this paper allows relaxing such an assumption, and provides a method to develop models that do not rely on a particular demand distribution assumption.

Dynamic Pricing and Revenue Management: Ambiguous Demand. Studies such as Aviv and Pazgal (2005) (and the references therein) develop POMDP models for dynamic pricing problems in revenue management. Again, using an APOMDP model instead of a POMDP allows a DM to reduce the dependency of the model to a specific demand distribution assumption, and thereby adopt policies that are robustly effective.

Optimal Search Problems: Ambiguous Moves. Search for a moving object is an important problem with various applications in national security and related fields. Pollock (1970) studies a model in which the target can be in any of two possible location areas. A hidden Markov chain that is assumed to be known by a searcher describes the movement of the object. The searcher can investigate either one of the two areas just before each transition of the object, which will result in conditional probabilities of the object being in the area searched. The goal is to find a strategy that will minimize the expected number of steps needed to find the object and the strategy that will maximize the detection probability in a given number of steps. In such search models, it is typically assumed that the movements of the object are probabilistically known to the searcher. However, in most applications the searcher does not have any way of determining such probabilities, especially since the object is not observable. Hence, the searcher must take into account the inevitable ambiguity with respect to the transition probabilities used. APOMDPs provide a natural tool to improve such models. Specifically, the search policy obtained by solving a particular POMDP depends heavily on the assumed dynamics of the underlying hidden Markov process. An APOMDP can reduce such a dependency and provide robust search policies.

Sequential Design of Experiments, Bayesian Control Models, and Bandit Problems. Various papers including Rieder (1991) and Krishnamurthy and Wahlberg (2009) discuss the connection between POMDPs, the sequential design of experiments, bandit problems, and more generally the class of Bayesian control models. APOMDPs can be used for such applications as well to take into account the inevitable model misidentifications, and reduce the dependency of actions to unknown probability measures.

Dynamic Principal-Agent Models. Dynamic principal-agent models are widely used in economics and operations management, among others. In particular, in studies such as Zhang and Zenios (2008) and Saghafian and Chao (2014), POMDP-type models are developed to address the

underlying information asymmetry and/or moral hazard aspects. It would be fruitful to observe the impact of APOMDPs on such classes of problems.

7. Connection to Stochastic Games with Perfect Information

We now briefly discuss an interesting connection between APOMDPs and nonzero-sum sequential games with perfect information and an uncountable state space. Consider a game with two players, player 1 (the DM) and player 2 who has two types: type 1 (adversary) and type 2 (ally). In state $\pi \in \Pi_{\mathscr{S}}$, player 1 chooses an action $a \in \mathscr{A}$ and receives a reward of $\pi' g^a$. Simultaneously, a biased coin that has probability $\alpha \in \mathscr{I}$ of yielding head is tossed. The state of the system becomes (π, a, ω), where $\omega \in \Omega \triangleq \{H, T\}$ is the outcome of the coin toss. If the outcome is head (tail), a type 1 (type 2) player plays (determines the model that dictates the observation and state transition probabilities). Consequently, the DM receives a signal/observation, and the new state becomes π', but this does not result in any reward for the decision maker. Each two sequential stages in this stochastic game correspond to one period in the APOMDP ($(2t, 2t+1)$ can be used to denote the stages of the stochastic game corresponding to period t of the APOMDP).

We note that, although the APOMDP is a sequential decision-making processes with imperfect information, the corresponding game is of perfect information[21]. To observe this, fix the pessimism factor $\alpha \in \mathscr{I}$ and define the game's state space as $\bar{\Pi}_{\mathscr{S}} = \Pi_{\mathscr{S}} \cup (\Pi_{\mathscr{S}} \times \mathscr{A} \times \Omega)$. For all $\pi \in \Pi_{\mathscr{S}} \subset \bar{\Pi}_{\mathscr{S}}$, let player 1 action space be $\mathscr{A}$ and that of player 2 be an arbitrary singleton. For all $\pi \in \Pi_{\mathscr{S}} \times \mathscr{A} \times \Omega \subset \bar{\Pi}_{\mathscr{S}}$, let player 1 action be an arbitrary singleton and that of player 2 be $m \in \mathscr{M}$. This shows that the game has a perfect information (with an uncountable state space).

Since the game is not necessarily zero-sum (unless $\alpha = 1$), the game falls within the class of sequential non-zero-sum games with perfect information and an uncountable state space. We refer to Whitt (1980), Nowak (1985), Nowak and Szajowski (1999), Simon (2007), and the references therein for some technical results on such games. Since the literature on these games is still limited and many technical challenges remain unsolved, we leave it for future research to use the above-mentioned link between APOMDPs and stochastic games to generate further structural results for APOMDPs.

8. Concluding Remarks

Motivated by various real-world applications, we develop a new framework for dynamic stochastic decision-making termed APOMDP, which allows for both incomplete information regarding the system's state and ambiguity regarding the correct model. The proposed framework is a generalization of both POMDPs and robust MDPs, in that the former does not allow for model misspecification,

[21] A two-player stochastic game is said to have perfect information if its state space can be partitioned into two subsets such that the action set for player 1 is a singleton on one partition and the action set for player 2 is a singleton on the other partition (see, e.g., p. 72 of Fudenberg and Tirole (1991), or p. 275 of Iyengar (2005)).

and the latter does not allow for incomplete state information. In addition, unlike the literature on robust MDP studies, the proposed approach in this paper considers a combination of worst and best outcomes (α-MEU preferences) with a controllable level of pessimism. This (a) results in a differentiation between ambiguity and ambiguity attitude, (b) avoids the over-conservativeness of traditional maximin approaches widely used in robust optimization approaches, and (c) is found to be suitable in laboratory experiments in various choice behaviors including portfolio optimization (as well as several other empirical studies that find that the inclusion of ambiguity seeking features is behaviorally meaningful). The α-MEU preferences also do not add much to the high computational complexity of dynamic models under incomplete information, especially in comparison to other preferences that may require consideration of all ambiguous outcomes (and not only the best and the worst).

To facilitate the search for optimal policies, we present several structural properties for APOMDPs. We find that model ambiguity in APOMDPs may result in non-convexity of the value function, hence deviating from the seminal result of Sondik (1971), who established the convexity of POMDP value functions (in finite-horizon settings). However, we present conditions under which this convexity result can be extended from POMDPs to APOMDPs. We do this by using the Blackwell Ordering (Blackwell (1951a)) and a variation of Blackwell-Sherman-Stein sufficiency theorem (Blackwell (1951a, 1953, 1951b), Stein (1951)) to connect the required condition to the notion of "model informativeness" in the "cloud" of models considered by the DM. We also briefly connect our result to a different way of handling model misspecification appeared in studies such as Hansen and Sargent (2007), in which beliefs are distorted (due to model ambiguity) using a martingale process. In addition to the value function, we also presented conditions for policy regions to be convex. These convexity results can significantly simplify the search for optimal policies in APOMDPs.

Using the TP_2 stochastic ordering, we present conditions under which monotonicity is preserved under both pessimism and optimism, and hence under the APOMDP Bellman operator. We also provide a performance guarantee for the maximum reward loss of a DM who uses the proposed APOMDP approach compared to an imaginary DM who does not have any model ambiguity. We generate further insights into the benefit and robustness of the proposed APOMDP approach through a numerical experiment. We show that, if hypothetically the DM is allowed to optimize his pessimism level, he would not chose extreme values corresponding to maximax or maximin preferences. Furthermore, using the Hausdorff distance, we show that policies adopted by the DM are not highly sensitive to his pessimism level: they remain the same for close pessimism levels.

We also discuss various applications of APOMDPs. For the class of machine replacement problems, we show how our structural results can help to establish the existence of control-limit policies, and even effectively approximate the control-limit thresholds. For several other applications, we

briefly discus why developing an APOMDP model can help to provide robust policies, but we leave it to future research to follow the use of APOMDPs in such applications. In light of our promising findings for policies obtained via APOMDPs, this can provide an influential path for future research.

Future research can also develop approximations, bounds, myopic, or other suboptimal policies for APOMDPs to further facilitate solving them. Another fruitful area of research is to use and advance results in non-zero sum stochastic games with uncountable state spaces to generate more insights into the structure of APOMDPs. Given various applications of APOMDPs including those briefly discussed in this paper, we expect to see more results from future research in such directions.

Acknowledgment. The author is grateful to *Hao Zhang* (Sauder School of Business, University of British Columbia) for his suggestions, comments, and discussions which helped to improve this paper.

References

Ahn, D., S. Choi, D. Gale, S. Kariv. 2007. Estimating ambiguity aversion in a portfolio choice experiment. Working Paper, UC Berkeley.

Aoki, M. 1965. Optimal control of partially observable Markovian systems. *Journal of Franklin Inst.* **280** 367–386.

Arrow, K. J., L. Hurwicz. 1997. An optimality criterion for decision making under ignorance. K. J. Arrow, L. Hurwicz, eds., *Studies in Resource Allocation Processes*. Cambridge University Press.

Arrow, K.J. 1951. Alternative approaches to the theory of choice in risk-taking situations. *Econometrica* **19**(4) 404–437.

Astrom, K.J. 1965. Optimal control of Markov decision processes with incomplete state estimation. *J. Math. Anal. Appl.* **10** 174–205.

Aviv, Y., A. Pazgal. 2005. A partially observed Markov decision process for dynamic pricing. *Management Science* **51**(9) 1400–1416.

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

Bertsekas, D. 1976. *Dynamic Programming and Stochastic Control*. Academic Press, New York.

Bertsekas, D.P., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Blemont, MA.

Bhidé, A.V. 2000. *The Origin and Evolution of New Business*. Oxford University Press, Oxford.

Blackwell, D. 1951a. Comparison of experiments. *2nd Berkeley Syposium on Mathematical Statistics and Probability*. University fo California Press, 93–102.

Blackwell, D. 1951b. Comparison of experiments. *Proceedings of the National Academy of Science*, vol. 37. 826–831.

Blackwell, D. 1953. Equivalent comparison of experiments. *Annals of Mathemtical Statistics* **24** 265–272.

Cassandra, A. R., L. P. Kaelbling, M. L. Littman. 1994. Acting optimally in partially observable stochastic domains. *Proceeding of the 12th National Conference of Artificial Intelligence (AAAI-94)* **1**(1) 1023–1028.

Cheng, H. T. 1988. Algorithms for partially observable Markov decision processes. Ph.D. thesis, University of British Columbia.

Chhatwal, J., O. Alagoz, E. S. Burnside. 2010. Optimal breast biopsy decision making based on mammographic features and demographic factors. *Operations Research* **58**(6) 1577–1591.

de Farias, DP, B. Van Roy. 2003. The linear programming approach to approximate dynamic programming. *Operations Research* **51**(6) 850–865.

DeHoratius, N., A.J. Mersereau, L Schrage. 2008. Retail inventory management when records are inaccurate. *Manufacturing and Service Operations Management* **10**(2) 257–277.

Delage, E., S. Mannor. 2010. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research* **58**(1) 203–213.

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

Eckles, J. E. 1968. Optimum maintenance with incomplete information. *Operations Research* **16**(1) 1058–1067.

Ehrenfeld, S. 1976. On a sequential Markovian decision procedure with incomplete information. *Computers and Operations Research* **3**(1) 39–48.

Epstein, L.G., M. Schneider. 2003. Recursive multiple priors. *Journal of Economic Theory* **113**(1) 1–31.

Feng, Z., S. Zilberstein. 2004. Region-based incremental pruning for POMDPs. *Proceeding of the 20th Conference of Uncertainty in Artificial Intelligence (UAI-04)* **1**(1) 146–153.

Fudenberg, D., J. Tirole. 1991. *Game Theory*. The MIT Press, Cambridge, MA.

Ghiradato, P., F. Maccheroni, M. Marinacci. 2004. Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory* **118** 133–173.

Ghirardato, P., F. Maccheroni, M. Marinacci. 2008. Revealed ambiguity and its consequences: updating. *Advances in decision making Under Risk and Uncertainty*. Springer, 3–18.

Grosfeld-Nir, A. 1996. A two-state partially observable Markov decision process with uniformly distributed observations. *Operations Research* **44**(3) 458–463.

Grosfeld-Nir, A. 2007. Control limits for two-state partially observable Markov decision processes. *European Journal of Operational Research* **182**(1) 300–304.

Hansen, L.P., T.J. Sargent. 2007. Recursive robust estimation and control without commitment. *Journal of Economic Theory* **136** 1–27.

Hansen, L.P., T.J. Sargent. 2008. *Robustness*. Princeton University Press, Princeton, NJ.

Hansen, L.P., T.J. Sargent. 2012. Three types of ambiguity. *Journal of Monetary Economics* **59** 422–445.

Heath, C., A. Tversky. 1991. Preference and belief: ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty* **4**(1) 5–28.

Hurwicz, L. 1951a. Optimality criteria for decision making under ignorance. *Cowles Commission discussion paper: Statistics no. 370* .

Hurwicz, L. 1951b. Some specification problems and applications to econometric models. *Econometrica* **19** 343–344.

Itoh, H., K. Nakamura. 2007. Partially observable Markov decision processes with imprecise parameters. *Artifical Inteligence* **171** 453–490.

Iyengar, G. N. 2005. Robust dynamic programming. *Mathematics of Operations Research* **30**(2) 257–280.

Jin, L., K. Kumagai, K. Suzuki. 2011. Control limit policy for partially observable Markov decision process based on stochastic increasing ordering. *Quality Technology & Quantitative Management* **8**(4) 479–493.

Karlin, S., Y. Rinott. 1980. Classes of orderings of measures and related correlation inequalities: I. Multivariate totally positive distributions. *J. Multivariate Analysis* **10** 467–498.

Klibanoff, P., M. Marinacci, S. Mukerji. 2005. A smooth model of decision making under ambiguity. *Econometrica* **73**(6) 1849–1892.

Klibanoff, P., M. Marinacci, S. Mukerji. 2009. Recursive smooth ambiguity preferences. *Journal of Econonimc Theory* **144** 930–976.

Krishnamurthy, V. 2011. Bayesian sequential detection with phase-distributed change time and nonlinear penalty a POMDP lattice programming approach. *IEEE Transactions on Information Theory* **57** 7096 – 7124.

Krishnamurthy, V., V. Djonin. 2009. Optimal threshold policies for multivariate POMDPs in radar resource management. *IEEE Transactions on Signal Processing* **57**(10) 3954–3969.

Krishnamurthy, V., B. Wahlberg. 2009. Partially observed Markov decision process multiarmed bandits structural results. *Mathematics of Operations Research* **34**(2) 287–302.

Lovejoy, W. S. 1987a. On the convexity of policy regions in partially observed systems. *Operations Research* **35**(4) 619–621.

Lovejoy, W. S. 1987b. Some monotonicity results for partially observed Markov decision processes. *Operations Research* **35**(5) 736–743.

Lovejoy, W. S. 1991a. Computationally feasible bounds for partially observed Markov decision processes. *Operations Research* **39**(1) 162–175.

Lovejoy, W. S. 1991b. A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research* **28**(1) 47–66.

Lovejoy, W. S. 1996. Comparison of some suboptimal control policies in medical drug therapy. *Operations Research* **44**(5) 696–709.

Lusena, C., J. Goldsmith, M. Mundhenk. 2001. Nonapproximability results for partially observable Markov decision processes. *J. Artificial Intelligence Res.* **14** 83–103.

Maillart, L. M., J. S. Ivy, S. Ransom, K. Diehl. 2008. Assessing dynamic breast cancer screening policies. *Operations Research* **56**(6) 1411–1427.

Maillart, L.M. 2006. Maintenance policies for systems with condition monitoring and obvious failures. *IIE Transactions* **38**(6) 463–475.

Marinacci, M. 2002. Probabilistic sophistication and multiple priors. *Econometrica* **70**(2) 755–764.

Marschak, J., K. Miyasawa. 1968. Economic comparability of information systems. *International Economics Review* **9** 137–174.

Monahan, G. E. 1982. A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science* **28**(1) 1–16.

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

Nowak, A.S. 1985. Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space. *Journal of Optimization Theory and Applications* **45**(4) 591–602.

Nowak, A.S., K. Szajowski. 1999. Nonzero-sum stochastic games. M. Bardi, T.E.S. Raghavan, T. Parthasarathy, eds., *Stochastic and Differential Games*, *Annals of the International Society of Dynamic Games*, vol. 4. Birkhuser Boston, 297–342.

Papadimitriou, C.H., J.N. Tsitsiklis. 1987. The complexity of Markov decision processes. *Mathematics of Operations Research* **12**(3) 441–450.

Peek, N. B. 1999. Explicit temporal models for decision-theoretic planning of clinical management. *Artificial Intelligence in Medicine* **15**(1) 135–154.

Perakis, G., G. Roels. 2008. Regret in the newsvendor model with partial information. *Operations Research* **56**(1) 188–203.

Pollock, S.M. 1970. A simple model of search for a moving target. *Operations Research* **18**(5) 883–903.

Poupart, P., C. Boutilier. 2004. Bounded finite state controllers. *Advances in Neural Information Processing Systems*. MIT Press, 823–830.

Rieder, U. 1991. Structural results for partially observed control models. *Methods and Models of Operations Research* **35** 473–490.

Ross, S. 1971. Quality control under Markovian deterioration. *Management Science* **17**(1) 587–596.

Saghafian, S., X. Chao. 2014. The impact of operational decisions on the optimal design of salesforce incentives. *Naval Research Logistics* **61**(4) 320–340.

Saghafian, S., B. Tomlin. 2015. The newsvendor under demand ambiguity: Combining data with moment and tail information. Working Paper, Arizona State University.

Shaked, M., J.G. Shanthikumar. 2007. *Stochastic Orders*. Springer, New York, NY.

Si, J., A.G. Barto, W.B. Powell, D. Wunsch. 2004. *Handbook of Learning and Approximate Dynamic Programming*. Wiley-IEEE Press.

Simon, R.S. 2007. The structure of non-zero-sum stochastic games. *Advances in Applied Mathematics* **38**(1) 1–26.

Smallwood, R., E. J. Sondik. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research* **21**(5) 1071–1088.

Smith, J.E., K.F. McCardle. 2002. Structural properties of stochastic dynamic programs. *Operations Research* **50**(5) 796–809.

Sondik, E. 1971. The optimal control of partially observable Markov processes. Unpublished Ph.D. dissertation, Stanford University, 1971.

Sondik, E. J. 1978. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations Research* **26**(2) 282–304.

Stein, C. 1951. Notes on a seminar on a theoritical statistics; comparison of experiments. Unpublished report.

Stoy, J. 2011. Statistical decisions under ambiguity. *Theory and Decision* **70**(2) 129–148.

Topkis, D.M. 1998. *Supermodularity and Complementarity*. Princton University Press, Princton, NJ.

Veatch, M.H. 2013. Approximate linear programming for average cost MDPs. *Mathematics of Operations Research* **38**(3) 535–544.

Wang, R. 1977. Optimal replacement policy under unobservable states. *Journal of Applied Probability* **14**(1) 340–348.

White, C.C. 1977. A Markov quality control process subject to partial observation. *Management Science* **23**(1) 843–852.

White, C.C. 1978. Optimal Inspection and Repair of a Production Process Subject to Deterioration. *Journal of the Operational Research Society* **29**(3) 235–243.

White, C.C. 1979. Optimal control-limit strategies for a partially observed replacement problem. *International Journal of Systems Science* **10**(1) 321–331.

Whitt, W. 1980. Representation and approximation of noncooperative sequential games. *SIAM Journal of Control and Optimization* **18**(1) 33–48.

Whitt, W. 1982. Multivariate monotone likelihood ratio and uniform conditional stochastic order. *Journal of Applied Probability* **19**(3) 695–701.

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

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

Zhang, H. 2010. Partially observable Markov decision processes: A geometric technique and analysis. *Operations Research* **58**(1) 214–228.

Zhang, H., S.A. Zenios. 2008. A dynamic principal-agent model with hidden information: Sequential optimality through truthful state revelation. *Operations Research* **56** 371–386.

Zhang, J., B.T. Denton, H. Balasubramanian, N.D. Shah, B.A. Inman. 2012. Optimization of prostate biopsy referral decisions. *Manufacturing and Service Operations Management* **14**(4) 529–547.

Zhang, N. L., W. Liu. 1996. Planning in stochastic domains: Problem characteristics and approximation. *Technical Report HKUST-CS96-31, Hong Kong University of Science and Technology* **1**(1) 1.