Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 178 (2018) 1–35

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Ambiguous partially observable Markov decision processes: Structural results and applications ☆

Soroush Saghafian

Harvard Kennedy School, Harvard University, Cambridge, MA, United States of America

Received 1 September 2016; final version received 2 March 2018; accepted 15 August 2018
Available online 20 August 2018

Abstract

Markov Decision Processes (MDPs) have been widely used as invaluable tools in dynamic decision-making, which is a central concern for economic agents operating at both the micro and macro levels. Often the decision maker's information about the state is incomplete; hence, the generalization to Partially Observable MDPs (POMDPs). Unfortunately, POMDPs may require a large state and/or action space, creating the well-known "curse of dimensionality." However, recent computational contributions and blindingly fast computers have helped to dispel this curse. This paper introduces and addresses a second curse termed "curse of ambiguity," which refers to the fact that the exact transition probabilities are often hard to quantify, and are rather ambiguous. For instance, for a monetary authority concerned with dynamically setting the inflation rate so as to control the unemployment, the dynamics of unemployment rate under any given inflation rate is often ambiguous. Similarly, in worker-job matching, the dynamics of worker-job match/proficiency level is typically ambiguous. This paper addresses the "curse of ambiguity" by developing a generalization of POMDPs termed Ambiguous POMDPs (APOMDPs), which 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 of transitions.

Importantly, this paper extends various structural results from POMDPs to APOMDPs. These results enable the decision maker to make robust decisions. 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, and (c) is found to be suitable in laboratory exper-

☆ The author is grateful (in no particular order) to Tomasz Strzalecki (Harvard), Richard Zeckhauser (Harvard), Maciej Kotowski (Harvard), Michael Veatch (Gordon College), and Hao Zhang (University of British Columbia) for valuable suggestions, comments, and discussions which helped to improve this paper. The author also thanks the editors, and the anonymous referee for their various helpful comments. This work was partially supported by the National Science Foundation under the Award Number CMMI-1562645.

E-mail address: soroush_saghafian@hks.harvard.edu.

https://doi.org/10.1016/j.jet.2018.08.006

iments 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. The analysis also identifies a performance guarantee for the proposed approach by developing a bound for its maximum reward loss due to model ambiguity.

JEL classification: C61; D81; D83; D84

Keywords: POMDP; Unknown probabilities; Model ambiguity; Structural results; Control-limit policies

1. Introduction

A critical factor for economic agents operating at both the micro and macro levels is decision-making in dynamic environments. Markov Decision Processes (MDPs) have been widely used for dynamic decision-making in such environments when two main assumptions hold: (1) the state of the system is completely known/observable at each decision epoch, and (2) the (Markovian) 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 real-world settings.

In such settings, 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 assumption (2) above. Hence, this paper extends POMDPs 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).*[1]

To address the curse of ambiguity, we assume that the decision maker simultaneously faces (a) non-probabilistic ambiguity (a.k.a. *Knightian uncertainty*) about the true model, and (b) probabilistic uncertainty or risk given the true model.[2] 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 this paper's framework, the decision maker is faced with Knightian uncertainty regarding the true model, while under each potential model, he has a certain probabilistic understanding of how observations and the core system state evolve over time. This draws a line between *ambiguity* (lack of

[1] To highlight the importance of considering the "curse of ambiguity," we note that the work of Savage and the applied statistical decision theory literature, which has been embraced by rational economists, suggests that probabilities should simply be estimated and that there should be no discount for ambiguity. However, the literature starting with Knight, and then dealing with the Ellsberg Paradox, and exploding on the scene with the work of Tversky and Kahneman recognizes that ambiguity plays an essential role in human decision-making.

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

knowledge about the true probability model) and *risk* (probabilistic consequences of decisions under a known probability model).

Another important element 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 (his desire level for ambiguity). Given an ambiguity set, the *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 (for related discussions, see, e.g., Ghirardato et al., 2004, Delage and Mannor, 2010, Xu and Mannor, 2012, Saghafian and Tomlin, 2016, and Bren and Saghafian, 2016). 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 which reveals 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 scenarios.

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 MEU approach and allow the DM to take into account both the worst possible outcome and the best possible outcome. The preferences under this criterion are called α-MEU preferences (with "multiple-priors"), and are axiomatized in Ghirardato et al. (2004) (see also Marinacci, 2002). They are found to be suitable for modeling various choice behaviors including those in portfolio selection (see, e.g., Ahn et al., 2007).

The key 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 discussed four axioms that a choice operator must follow, and demonstrated that under complete ignorance, one can restrict attention merely to the extreme outcomes (i.e., the best and the worst). The work of Hurwitcz and Arrow in early 1950s constructed a collection of utility functions for a DM under ambiguity including a convex combination 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 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 studies above where it is constrained to be one. 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, unlike the above-mentioned studies, our work is also applicable in several applications where the state is hidden to the decision maker (for some examples in the economics literature, see, e.g., Jovanovic, 1979, 1982, Jovanovic and Nyarko, 1995, 1996, Hansen and Sargent, 2007, and Cogly et al., 2008). 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.[3]

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 MDPs 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, 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 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 complete metric space) and convergence of a finite-horizon setting to that of an infinite-horizon, we show that unlike the seminal result of Smallwood and Sondik (1973) (see also Sondik, 1971 and Sondik, 1978) who proved the convexity (and piecewise-linearity) of the value function for POMDPs, 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 Smallwood and Sondik (1973) 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 already developed in the literature for POMDPs. Furthermore, using the Blackwell ordering (Blackwell, 1951a), which is often referred to as *information garbling* in the economics of information literature (see, e.g., Marschak and Miyasawa, 1968), 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 (see, e.g., Hansen and Sargent, 2007).

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 experiment, we then show that the APOMDP approach is indeed robust to model misspecification. More importantly, we show that

[3] We will discuss a variety of such applications from economics and beyond in Sections 6 and 8.

the proposed APOMDP approach provides more effective policies than those provided by traditional maximax or maximin criteria. 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 from economics and beyond. 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 in a variety of applications. This is achieved by reducing the reliance on a specific probabilistic model.

In particular, while some of our numerical experiments briefly illustrate the application of our APOMDP framework in dynamically adjusting inflation rate so as to control unemployment level, to further illustrate the advantages of the meta-structural results provided in the paper, we discuss two specific applications of APOMDPs in more detail. The first application is in job matching models. We extend the literature on such models by allowing for model ambiguity. Specifically, we consider the discrete-time version of the well-known job matching model of Jovanovic (1979) (see also Sections 10.10 and 10.11 of Stokey et al., 1989), and provide an extension by considering the fact that the dynamics of worker-job match level often cannot be quantified via a single probabilistic model. We discuss how this extension can be modeled as an APOMDP, and how the structural results developed for general APOMDPs significantly simplify the complexity of identifying the optimal policy. We also show how the performance guarantee developed for general APOMDPs can be used to quantify a *price of ambiguity* in job matching problems with model ambiguity. The second application that we use to illustrate the advantage of our meta-structural results is the class of machine replacement problems which is concerned with optimal timing of replacing a general asset ("machine" hereafter); see, e.g., Cooper and Haltiwanger (1993) and the references therein for applications of this class of problems in economic theory. 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 this reality. Furthermore, based on the general structural properties established for APOMDPs, we 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 threshold. In addition to these applications, we conclude this paper by shedding light on a variety of other applications where APOMDPs can be remarkably useful. These include areas such as strategic pricing, dynamic principal-agent models, inventory control, optimal search, medical decision-making,[4] sequential design of experiments, Bayesian control, and bandit problems.

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.

[4] We have specifically observed the benefits of using the APOMDP framework proposed in this paper in a real-world medical decision-making problem faced by physicians in the Mayo Clinic (see, e.g., Boloori et al., 2018 for more details).

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 specific applications of APOMDPs, 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, and Online Appendix B discusses two ways to guarantee dynamic consistency of preferences in the proposed APOMDP framework.

2. Literature review

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). Dynamic decision-making under ambiguity has been studied under maximin expected utility with multiple-priors (see, e.g., Gilboa and Schmeidler, 1989 and Epstein and Schneider, 2003), multiplier preferences (see, e.g., Hansen and Sargent, 2001 and Strzalecki, 2011), and variational preferences (see, e.g., Maccheroni et al., 2006). When preferences are dynamically inconsistent, Siniscalchi (2011) provides a unique and in-depth decision-theoretic framework for studying dynamic choice.

The α-MEU preferences that we use in this paper is discussed and axiomatized in Marinacci (2002) and Ghirardato 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,[5] especially if used for POMDPs which are already computationally very complex. Since the α-MEU criterion only requires considering two ambiguous outcomes—the best and the worst—the additional computational burden is not as significant.

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 economics, operations research, medicine, biology, computer science, and engineering, among others, and (b) the fact that in most applications, model parameters cannot be exactly estimated (e.g., 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), Hansen and Sargent (2007), and Osogami (2015). However, these papers differ from our work in two main ways: (a) they do not develop detail meta-structural results (e.g., convexity,

[5] The computational burden is lower in repeated but not fully dynamic decision-making settings; see, e.g., Saghafian and Tomlin (2016) for an application in a repeated newsvendor setting.

monotonicity, etc.) that can simplify the search for optimal policies in various applications 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 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 through learning. For this stream of research, we refer interested readers to Saghafian and Tomlin (2016) (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. Using incoming data to learn about (and overcome) model ambiguities in a POMDP framework has also appeared in Bren and Saghafian (2016), where a specific decision-making context—control of multi-class queueing systems—is considered. Unlike such studies, and similar to the literature on robust MDPs, the goal in this paper is not to reduce or overcome ambiguity using incoming data (e.g., through learning). 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 $(\alpha, \beta, \mathscr{S}, \mathscr{O}, \mathscr{A}, \mathscr{G}, \mathscr{P}, \mathscr{R})$. 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* which represent the sets of possible transition probability matrices with respect to core states and observations, respectively.[6,7]

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.[8] 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

[6] 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).

[7] For general information regarding use and construction of ambiguity sets, we refer interested readers to studies such as Gupta (2018) and the references therein. Particularly, we note that there exist relatively standard methods for constructing the ambiguity sets with respect to transition probabilities using data and/or expert opinion (see, e.g., Wiesemann et al., 2013, Nilim and El Ghaoui, 2005, Saghafian and Tomlin, 2016, and the reference therein). For instance, one can use data along with the Baum–Welch algorithm (see, e.g., Welch, 2003) to create point estimates for state and observation transition probabilities, and use such standard methods to create ambiguity sets around the point estimates. Using expert opinion (see, e.g., Saghafian and Tomlin, 2016) is another method to directly build the ambiguity sets. In Remark 3, we also briefly discuss the idea of using martingale distortions to indirectly build the ambiguity sets. However, to be general, we do not restrict our approach to a specific method of constructing the ambiguity sets.

[8] 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 structural results in this paper can be 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. However, increasing the size of $\mathscr{M}$ may affect the achievable performance guarantee; see, e.g., Corollary 1.

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 throughout the paper that $\alpha \in \mathscr{I}$ and $\beta \in [0, 1)$.

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 $\boldsymbol{\pi} \in \Pi_\mathscr{S}$ could be obtained by a traditional POMDP Bellman equation (along with the terminal condition $V_0^m(\boldsymbol{\pi}) = \boldsymbol{\pi}' g_0$ for some $g_0 \in \mathbb{R}^n$):

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

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, $Pr\{o|\boldsymbol{\pi},a,m\} = \sum_i \sum_j \boldsymbol{\pi}_i p_{ij}^a(m) r_{jo}^a(m)$ is the probability of observing o under belief $\boldsymbol{\pi}$, 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):

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

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

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

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

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

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[9]—knows m: he only knows that $m \in \mathscr{M}$. Hence, the reward and policy cannot be simply obtained from (4). How can the DM derive a policy then that is optimal without adding much to the underlying computational complexity?

To answer this question, 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 $(t-1)$-period's expected reward-to-go function, respectively:

[9] If it was probabilistically known, then the DM could form a prior about the true model and include that in the state space. This would transfer the problem back to a traditional POMDP (with an augmented state space) in which observations are used over time to learn about the true model.

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

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

In this setting,

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

and $V_{t-1}(\boldsymbol{\pi}, \alpha)$ denotes the decision maker's reward with $t-1$ periods to go.[10] Using this notation, the DM can derive a policy by solving the following dynamic program[11] (along with the terminal condition $V_0(\boldsymbol{\pi}, \alpha) = \boldsymbol{\pi}' g_0$ for some $g_0 \in \mathbb{R}^n$):

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

We refer to (8) as the finite-horizon APOMDP Bellman equation, and call the policy obtained by solving it α-Hurwicz[12] or H^α for short.[13] For notational convenience, we define the utility function

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

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

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

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 complete metric space $(\mathscr{B}^\alpha, d^\alpha)$ (see, e.g., Theorem 3.2 in Stokey et al. (1989) for a general contraction mapping result on complete metric spaces), where for any $V, W \in \mathscr{B}^\alpha$, the metric d^α is defined as $d^\alpha(V, W) \triangleq \sup_{\boldsymbol{\pi} \in \Pi_{\mathscr{S}}} |V(\boldsymbol{\pi}, \alpha) - W(\boldsymbol{\pi}, \alpha)|$. 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 space $(\mathscr{B}^\alpha, d^\alpha)$. That is, for any $V, W \in \mathscr{B}^\alpha$: $d^\alpha(\mathfrak{L}^\alpha W, \mathfrak{L}^\alpha V) \leq \beta\, d^\alpha(W, V)$.*

[10] Note that since the best and worst-case models are chosen independently of the previous periods best and worst-case model selections, our setting satisfies a similar requirement to the rectangularity assumption discussed by Epstein and Schneider (2003) in their recursive multiple-priors setting, and used by Nilim and El Ghaoui (2005) and Iyengar (2005) for robust MDPs. See also Riedel (2009) for related discussions on the use of the rectangularity assumption for optimal stopping problems with multiple-priors.

[11] For further discussion on the rationale for this dynamic program, see Remark 2 and Online Appendix B.

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

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

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 on the complete metric space $(\mathscr{B}^{\alpha}, d^{\alpha})$. To consider infinite-horizon APOMDPs, we let $T = \infty$ and denote the infinite-horizon APOMDP value function by $V_{\infty}(\boldsymbol{\pi}, \alpha)$.

Proposition 1 *(APOMDP convergence). For all $\pi \in \Pi_{\mathscr{S}}$ and $\alpha \in \mathscr{I}$, $V_{\infty}(\boldsymbol{\pi}, \alpha)$ is the unique solution to $\mathfrak{L}^{\alpha} V_{\infty}(\boldsymbol{\pi}, \alpha) = V_{\infty}(\boldsymbol{\pi}, \alpha)$. Furthermore, for all $\pi \in \Pi_{\mathscr{S}}$ and $\alpha \in \mathscr{I}$, $\lim_{t\to\infty} V_t(\boldsymbol{\pi}, \alpha) = V_{\infty}(\boldsymbol{\pi}, \alpha)$, 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 some dynamic settings might lead to violations of dynamic consistency in preferences. However, as we discuss in detail in Online Appendix B, dynamic consistency in our APOMDP 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 (see also footnote 10 for a related discussion on a special case with $\alpha = 1$ and fully observable core states). 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 adjustments to the pessimism level do not affect the adopted policy by the DM. Dynamic consistency provides a rationale to solve a multiple-period problem via dynamic programming. Nevertheless, dynamic programming can typically be used to derive effective policies. Indeed, it should be noted that although in Online Appendix B we discuss in detail two ways to guarantee dynamic consistency in our proposed APOMDP approach, the issue of dynamic consistency of preferences is not of first order importance in our work. The main reason is that, motivated by various applications (see, e.g., Section 6), our goal in this paper 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 Bellman equation 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. Hence, the policies obtained from solving our APOMDP program can be regarded as useful polices for various applications, even in special cases where the DM preferences do not exhibit dynamic consistency. Further discussions about dynamic consistency can be found in Online Appendix B.

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 Smallwood and Sondik (1973) who proved the convexity of the value function for POMDPs (in finite-horizon settings), we observe that the APOMDP value

function is not always convex in $\boldsymbol{\pi} \in \Pi_{\mathscr{S}}$. Hence, unlike POMDPs, the APOMDP value function does not always admit the desirable form $V_t(\boldsymbol{\pi}, \alpha) = \max_{\psi \in \Psi_{t,\alpha}} \{\boldsymbol{\pi}' \psi\}$ for some finite set of vectors $\Psi_{t,\alpha}$. We illustrate this through the following example.

Example 1 *(Non-convex value function).* Consider a representative APOMDP with $n = k = l = 3$ (i.e., three states, three observations, and three actions). For instance, the core states in this APOMDP may represent three levels of unemployment (e.g., low, medium, high), the observations may represent the imperfect signals that a monetary authority receives about the state of the unemployment being at any of the three levels, and the actions may represent three levels of inflation (e.g., low, medium, high) that can be targeted by the authority.

Suppose there are three models, $m = 1$, $m = 2$, and $m = 3$, with representative transition probabilities given for each model in a separate row below.

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

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

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

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

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

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

The existence of different models may represent the fact that the exact probabilistic dynamics of unemployment levels under any targeted inflation rate is not completely known, and is subject to misspecifications. Also, let the representative (scaled) rewards obtained under any state-action pair (e.g., unemployment-inflation level) be $g_0 = (1.0, 1.1, 1.0)'$, $g^1 = (1.5, 1.8, 1.6)'$, $g^2 = (1.7, 1.4, 1.5)'$, $g^3 = (1.6, 1.7, 1.5)'$, and assume $T = 3$, and $\beta = 0.9$.

Fig. 1 illustrates the value function for $\alpha = 0.95$ and $t = 3$ ($V_3(\boldsymbol{\pi}, 0.95)$ for various belief points $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3 = 1 - \pi_1 - \pi_2) \in \Pi_{\mathscr{S}}$). As Fig. 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 Smallwood and Sondik (1973) for POMDPs can be extended to APOMDPs.

Definition 1 *(Belief-Independent Worst-Case (BIWC) member).* The ambiguity set $\mathscr{M}$ is said to have a belief-independent worst-case (BIWC) member if $\underline{m}_t(\boldsymbol{\pi}, a, \alpha)$ is constant in $\boldsymbol{\pi}$ ($\forall t \in \mathscr{T}$, $\forall a \in \mathscr{A}$, $\forall \alpha \in \mathscr{I}$).

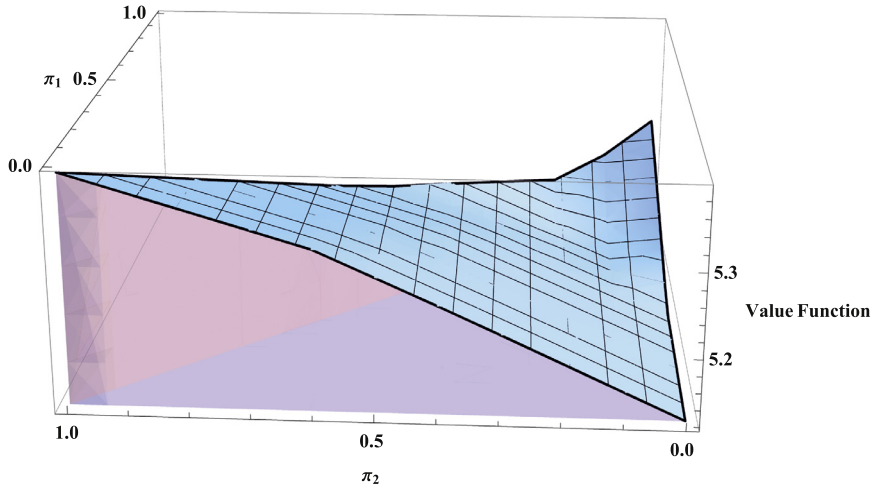


Fig. 1. The APOMDP value function of Example 1.

Proposition 2 *(Piecewise-linearity and convexity). If the ambiguity set $\mathscr{M}$ has a BIWC member, then $V_t(\boldsymbol{\pi}, \alpha)$ for any $t \in \mathscr{T}$ and $\alpha \in \mathscr{I}$ is piecewise-linear and convex in $\boldsymbol{\pi}$, and hence admits $V_t(\boldsymbol{\pi}, \alpha) = \max_{\psi \in \Psi_{t,\alpha}} \{\boldsymbol{\pi}' \psi\}$ for some finite set of vectors $\Psi_{t,\alpha}$.*

Proposition 2 extends the seminal result of Smallwood and Sondik (1973) from POMDPs to APOMDPs. The importance of this result is that it allows solving APOMDPs in a similar manner to POMDPs: one only needs to characterize the ψ-vectors introduced above, which are often referred to as α-vectors in the traditional POMDP literature (see, e.g., Monahan, 1982 for a review of related approaches).

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 BIWC member to be different in each period, or change based on the action or conservatism level. However, to guarantee a convex value function of $\boldsymbol{\pi}$ for *any*[14] value of conservatism level α and for the very general APOMDP approach introduced earlier, Proposition 2 requires the adversary (but not necessarily the ally) part of the nature to act independently of the DM's belief, $\boldsymbol{\pi}$. For instance, the DM may share its belief only with the ally part of the nature. Importantly, however, the ambiguity set can have a BIWC member under various situations, 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 and Sulganik, 2003), which is often referred to as *information garbling* in the economics of information literature (see, e.g., Marschak and Miyasawa, 1968). The property

[14] For special cases of α milder conditions are sufficient. For instance, with $\alpha = 0$, no condition is required: $V_t(\boldsymbol{\pi}, 0)$ is always convex in $\boldsymbol{\pi}$ even when $\mathscr{M}$ does not have a BIWC member.

above can be understood by noting that $P^a_m R^a_m$ 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^a_m) 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 these signals as outputs of statistical experiments under m^* and m, we first state the following variation of the Blackwell–Sherman–Stein sufficiency theorem[15] to connect model informativeness in our framework with convex stochastic ordering[16] (denoted by $\preceq_{cx}$) of the posterior likelihood distributions defined by the operator in (2). This will then allow us to connect model informativeness to the existence of a BIWC member.

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:*

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

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

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

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 BIWC 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 BIWC member. Furthermore, $\underline{m}_t(\boldsymbol{\pi}, a, \alpha) = m^*(a)$ 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, and hence, the ambiguity set will have a BIWC 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). In particular, (11) holds if, and only if, there are two random variables X and Y defined on the same probability space such that (1) $X =_{s.t.} T(\boldsymbol{\pi}, a, O, m_1)$ and $Y =_{s.t.} T(\boldsymbol{\pi}, a, O, m_2)$, and (2) $\{X, Y\}$ is a martingale (i.e., $\mathbb{E}[Y|X] = X$ a.s.). This means that (11) and

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

[16] For more details about convex stochastic ordering, see, e.g., Chapter 3 of Shaked and Shanthikumar (2007).

Proposition 3 can be viewed as follows. Suppose at each period, the DM uses a model $m^*(a)$ as his reference model under action $a \in \mathscr{A}$ to form a nominal posterior distribution over the hidden state (conditioned on his current belief). He then uses a martingale distortion of this posterior distribution as the set of possible posterior distributions, since he does not fully trust his reference model $m^*(a)$. That is, all his posterior distributions over the hidden state (i.e., under any $m \in \mathscr{M} \setminus m^*(a)$) represent martingale distorted versions of the posterior distribution formed under $m^*(a)$. 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 BIWC member, and most of our main results will still hold. This builds a subtle bridge between the "cloud" of models, $\mathscr{M}$, considered in this paper and the idea of using martingale distortions (without commitment to previous period distortions) to represent model misspecification that has appeared in Hansen and Sargent (2007).

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 $\Pi_{t,a}^*(\alpha) \triangleq \{\boldsymbol{\pi} \in \Pi_{\mathscr{S}} : a_t^*(\boldsymbol{\pi}, \alpha) = a\}$, which we refer to as *policy regions*. The search for an optimal policy can be significantly simplified if policy regions are convex.

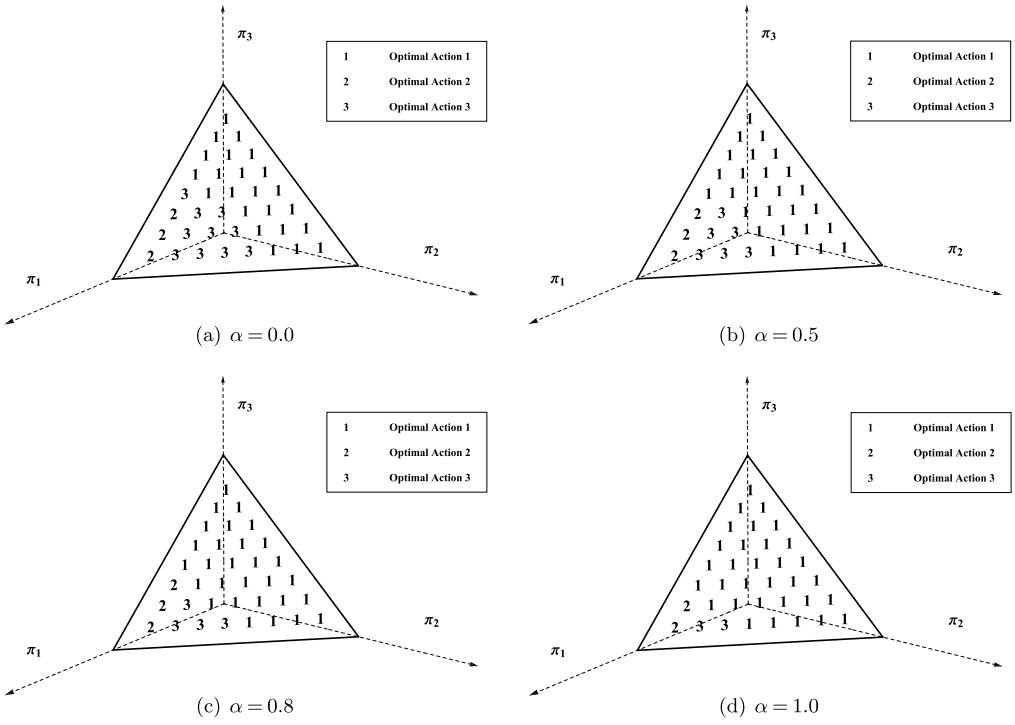
First, we note that even in a traditional POMDP, the policy regions may not be convex unless some specific conditions hold (see, e.g., Ross, 1971, White, 1978, Lovejoy, 1987a). We illustrate through the following example that the same observation holds for APOMDPs.

Example 2 *(Policy regions).* Consider the representative APOMDP of Example 1. Fig. 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 DM (Fig. 2 part (d)) will use action 1 (e.g., a low inflation target) unless he is somehow confident that the system is at state 1 (e.g., a low unemployment rate). Such a DM will use action 1 more than any other DM. In contrast, a maximax DM (Fig. 2 part (a)) uses action 3 (e.g., a high inflation target) more than any other DM. 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 BIWC member, and (b) under an action $a \in \mathscr{A}$, the value function $V_{t-1}(T(\boldsymbol{\pi}, a, o, m), \alpha)$ under both $\underline{m}_{t-1}$ and $\overline{m}_{t-1}$ is constant in $\boldsymbol{\pi}$, then $\Pi_{t,a}^*(\alpha)$ is a convex set ($\forall \alpha \in \mathscr{I}$).*

Condition (a) of the above proposition holds if, for instance, under each action 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 Section 6.1 we will show how the structural properties established in this section can be used to significantly reduce the complexity of solving an extension of job matching problems that, unlike the traditional literature, allows for model ambiguity. As we will see, for these problems, when the worker decides to switch jobs, his updated belief $T(\boldsymbol{\pi}, a, o, m)$ is independent of his prior belief $\boldsymbol{\pi}$ (although it depends on the underlying model, m). As another example that will discuss later, in machine replacement problems (see, e.g., Cooper and Haltiwanger, 1993 for introduction and applications in economic theory), machine (i.e., a general asset) deterioration is typically ambiguous and hard to define through one probabilistic model, making the proposed APOMDP framework



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

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

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

an attractive decision-making tool. However, when the machine is replaced or fully inspected, the system's core state becomes fully observable to the DM. Hence, under this action, $T(\boldsymbol{\pi}, a, o, m)$ lies on a corner of the $(n-1)$-simplex $\Pi_{\mathscr{S}}$ regardless of π. In Section 6.2, we will discuss this class of problems in more depth, and we will show how the structural properties established in this section can significantly simplify solving such a challenging class of problems.

Remark 4 *(APOMDP approximation).* As discussed earlier, the BIWC member condition required in several results in this section is not that restrictive, and holds under various conditions and in many applications. However, even for APOMDPs in which the ambiguity set does not have a BIWC member, our results can be used to provide effective approximations. For instance, one can enlarge the ambiguity set by adding fictitious members so that it satisfies the BIWC member requirement. Then, considering the APOMDP with the new ambiguity set as an approximation for the original APOMDP problem, one can use the results of this section to (a) approximate the value function with a piecewise-linear and convex function (Proposition 2), (b) calculate the approximated value function through characterizing the ψ-vectors (Proposition 2), (c) derive convex policy regions (Proposition 4), and (d) gain some insights into effective and yet well-structured control policies.

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. 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 $V_t(\boldsymbol{\pi}, \alpha)$ is non-increasing in α ($\forall t \in \mathscr{T}$, $\forall \boldsymbol{\pi} \in \Pi_{\mathscr{S}}$).*

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) $\boldsymbol{\pi} = (\boldsymbol{\pi}_i : i \in \mathscr{S}) \in \Pi_{\mathscr{S}}$ and $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\pi}}_i : i \in \mathscr{S}) \in \Pi_{\mathscr{S}}$, we use the notation $\boldsymbol{\pi} \preceq_{TP_2} \hat{\boldsymbol{\pi}}$, if $\boldsymbol{\pi}_i \hat{\boldsymbol{\pi}}_{\hat{i}} \geq \boldsymbol{\pi}_{\hat{i}} \hat{\boldsymbol{\pi}}_i$ 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.[17]

Definition 4 *(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 $a \in \mathscr{A}$.[18] 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 5 *(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 $f(\boldsymbol{\pi}, \ldots, .) \leq f(\hat{\boldsymbol{\pi}}, \ldots, .)$, whenever $\boldsymbol{\pi} \preceq_{TP_2} \hat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}, \hat{\boldsymbol{\pi}} \in \Pi_{\mathscr{S}}$.

In a POMDP, it is known that the value function belongs to $\mathscr{F}_{\Pi_{\mathscr{S}}}$ under some specific conditions (see, e.g., Proposition 1 of Lovejoy, 1987b, or Theorem 2.4 of Rieder, 1991). A natural question is whether or not the APOMDP value function belongs to $\mathscr{F}_{\Pi_{\mathscr{S}}}$. To provide an answer, we start with the following lemma, which shows that the set $\mathscr{F}_{\Pi_{\mathscr{S}}}$ is closed under both pessimism and optimism.

[17] A function f is said to be reflexive TP_2 (or, for simplicity, TP_2) if $f \preceq_{TP_2} f$.

[18] 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 $a \in \mathscr{A}$.

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

In Online Appendix A, we also establish the closedness of $\mathscr{F}_{\Pi_{\mathscr{S}}}$ under observation-based expectation operators (see Lemma EC.1). Using these results, we now first show that under some conditions the set $\mathscr{F}_{\Pi_{\mathscr{S}}}$ is closed under APOMDP value iteration. This, in turn, will allow 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.[19]

Proposition 6 *(Monotonicity preservation in APOMDP). Suppose the set of kernels P_m and R_m are TP_2 for all $m \in \mathscr{M}$.*
(i) If $V_{t-1}(\boldsymbol{\pi}, \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$, then $h_{t-1}(\boldsymbol{\pi}, a, m, \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ for all $m \in \mathscr{M}$.
(ii) If $V_{t-1}(\boldsymbol{\pi}, \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ and $g^a \in {}^{\uparrow}\mathbb{R}^n$ for all $a \in \mathscr{A}$, then $V_t(\boldsymbol{\pi}, \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$.

Finally, the following theorem presents conditions for both finite and infinite-horizon settings under which the value function of an APOMDP is monotonic. Notably, it provides a generalization for Proposition 1 of Lovejoy (1987b) and Theorem 4.2 of Rieder (1991) which establish monotonicity 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 in economics and beyond, 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 $V_t(\boldsymbol{\pi}, \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ for all $t \in \mathscr{T}$ and $\alpha \in \mathscr{I}$.
(ii) If $T = \infty$, then $V_\infty(\boldsymbol{\pi}, \alpha) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ 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 6 *(ϵ-Tightness).* The ambiguity set $\mathscr{M}$ is said to be ϵ-tight if for any two $m_1, m_2 \in \mathscr{M}$:

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

An APOMDP is said to be ϵ-tight if its ambiguity set is ϵ-tight.[20]

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

[20] 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.

The notion of ϵ-tightness defined above can also be viewed as a way of measuring the "diameter" of the ambiguity set $\mathscr{M}$. Using this notion, we now bound the maximum difference in the vector of conditional observation probabilities $Pr(o|\boldsymbol{\pi}, a, m) \triangleq (Pr\{o|\boldsymbol{\pi}, a, m_1\} : o \in \mathscr{O})$ caused by model ambiguity.

Lemma 4 *($\mathcal{L}_1$-norm bound). For any ϵ-tight APOMDP:*

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

where $||\cdot||_1$ is the $\mathcal{L}_1$-norm,[21] *$\xi = \min\{k\, n\, \epsilon, 2\}$, $n = |\mathscr{S}|$, 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 4, but by considering the case where the DM's belief state is distorted.

Lemma 5 *(Belief distortion). For any two belief states $\boldsymbol{\pi}, \hat{\boldsymbol{\pi}} \in \Pi_{\mathscr{S}}$:*

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

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

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

where ξ is defined in Lemma 4.

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 $V_t^{m,\eta}(\boldsymbol{\pi})$ the reward obtained under such policy when m is the true model and the belief is $\boldsymbol{\pi} \in \Pi_{\mathscr{S}}$. Assuming model m is the true model, we let $V_t^{m,\eta}(\hat{\boldsymbol{\pi}})$ denote the reward obtained under the same actions used to calculate $V_t^{m,\eta}(\boldsymbol{\pi})$, but when the belief is $\hat{\boldsymbol{\pi}} \in \Pi_{\mathscr{S}}$ instead of $\boldsymbol{\pi} \in \Pi_{\mathscr{S}}$ (i.e., a distorted belief).

Moreover, without loss of generality and for clarity, we assume $g_0 = 0$ in the rest of this section.[22]

Lemma 6 *(Max reward loss – distorted belief). Under any policy η:*

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

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

[21] For any real vector $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$ and real number $p \geq 1$, the $\mathcal{L}_p$-norm is $||x||_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$. In this paper, we use the $\mathcal{L}_1$-norm (i.e., the rectilinear distance) and the $\mathcal{L}_\infty$-norm (i.e., the limit of $\mathcal{L}_p$ as $p \to \infty$, or equivalently $||x||_\infty = \max\{|x_1|, |x_2|, \cdots, |x_n|\}$).

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

the answer, we let $V_t^{\eta}(\boldsymbol{\pi}, \alpha)$ denote the APOMDP value function under policy η, and compare it with the corresponding POMDP value function under η and model m, denoted by $V_t^{m,\eta}(\boldsymbol{\pi})$.

Lemma 7 *(Max reward loss – arbitrary policy). If the APOMDP is ϵ-tight, then under any policy η:*

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

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

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*

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

where $\bar{\xi}$ is defined in Lemma 7.

The bound provided in Theorem 2 is tight. For instance, it goes to zero as $\epsilon \to 0$. In various applications, this allows the DM to quantify a useful performance guarantee (see, e.g., Section 6.1 where we develop a "price of ambiguity" in a job matching application). Importantly, 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 representative APOMDP of Example 1, and to gain insights into the performance of the H^{α} policy, we consider different DMs (e.g., monetary authorities trying to dynamically set the inflation rate to control unemployment) and uniformly distribute them over the simplex $\Pi_{\mathscr{S}}$. The location of each DM represents his starting belief point (e.g., initial prior on the state of unemployment). 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

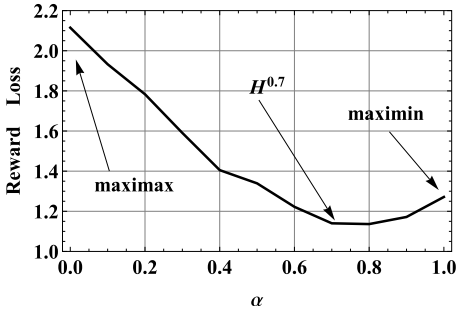


Fig. 3. The reward loss with model ambiguity under the H^α policy for various levels of α. Dynamic versions of widely used maximin ($\alpha = 1$) and maximax ($\alpha = 0$) policies are dominated by the H^α policy for some mid-level α (e.g., $\alpha \in [0.6, 0.9]$).

performance of $\binom{20+3-1}{3-1} = 231$ different DMs.[23] 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 as its average reward loss exactly (no approximation). Fig. 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.[24]

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 7 *(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

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

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

[24] 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.

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*

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

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 from economics (and beyond). As mentioned earlier, our work is widely applicable in various applications of dynamic decision-making where (a) the state is hidden to the decision maker (for some examples in the economics literature, see, e.g., Jovanovic, 1979, 1982, Jovanovic and Nyarko, 1995, 1996, Hansen and Sargent, 2007, and Cogly et al., 2008), and (b) the decision maker is facing model ambiguity/misspecification, and hence, wants to be robust to such ambiguity.

To illustrate the advantage of the results developed in previous sections, in what follows we first discuss an extension of the celebrated job matching model of Jovanovic (1979) in discrete time, where unlike the existing literature we allow for model ambiguity. As noted earlier, allowing for model ambiguity enables making robust decisions by reducing the reliance on a single probabilistic model. We show how this extension of the job matching problem can be modeled as an APOMDP, and how the structural results developed in the previous sections can be used to significantly reduce the complexity of characterizing the value function as well as the optimal policy. As another application, we next provide an extension for the widely studied class of machine/asset replacement problems, where unlike the extant literature we allow deterioration probabilities to be ambiguous (see, e.g., Cooper and Haltiwanger, 1993 for applications in settings without ambiguity in economic theory). Similar to the job matching problem, we then show how the structural results provided in the earlier sections help to simplify solving this challenging class of problems.

6.1. Job matching problems with model ambiguity

Models of job matching typically assume that a worker-task match can be fully defined via a single probabilistic model. Consider, for example, the following discrete-time version of the job matching model of Jovanovic (1979) (see also Sections 10.10 and 10.11 of Stokey et al., 1989). A worker can choose among different tasks, and each task has a worker-task match level, $\theta \in [0, 1]$, which represents the proficiency of the worker at that specific task. Suppose the worker-task match can be at one of the n different levels[25]: $\theta \in \{\theta_i : i = 1, 2, \cdots, n\}$. Without loss of generality, assume these levels are labeled such that $0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_n \leq 1$. The worker does not know his proficiency on a given task, and has to try it out, observe the returns,

[25] The number of levels, n, can be set to an arbitrarily high number to closely approximate models with continuous worker-task proficiency.

and form beliefs. At any given period, when working on a task with worker-task match θ_i, the worker produces a return of 1 with probability θ_i or a return of 0 with probability $1-\theta_i$, where returns on a given task are serially independent.[26] Upon observing the return at each period, the worker can decide to stay with the same job, or receive an average compensation, $c \in \mathbb{R}$ (with $c=0$ as a special case[27]), and draw a new θ in the next period (e.g., do a job search and engage in a new job with a potentially different worker-task match level).

The above setting represents a relatively general model of job matching in discrete time (especially in the spirit of Jovanovic, 1979). However, the literature on this type of models typically assumes that the worker can fully specify a single probabilistic model (e.g., a single probability mass function on θ denoted by $\mu(\theta)$) that defines for him the distribution of tasks among different worker-task match levels. This assumption is, however, strong and is unlikely to hold in realistic scenarios. What if the worker is ambiguous about this, and wants to take this ambiguity into account? That is, what if instead of a single probabilistic model he is facing a "cloud" of models? In particular, what if he cannot assign a single probability mass function μ on θ, and only knows that $\mu \in \{\mu_m : m = 1, 2, \cdots, |\mathscr{M}|\}$ for some $|\mathscr{M}| \in \mathbb{Z}_+$?[28]

In what follows, we show how the decision-making problem faced by the worker can be formulated as an APOMDP, and how our results in the previous section can be applied to derive the worker's optimal policy. To this end, similar to the general framework introduced in Section 3, let the core state space be $\mathscr{S} \triangleq \{1, \cdots, n\}$, where the core state represents the current task. That is, the core state is $i \in \mathscr{S}$ if the worker is engaged in a task with worker-task match θ_i. This core state is, however, hidden to the worker. Thus, we let $\boldsymbol{\pi} \in \Pi_{\mathscr{S}}$ denote his belief about the core state, where as before $\Pi_{\mathscr{S}}$ denotes the $(n-1)$-simplex induced by $\mathscr{S}$. By trying the task, the worker can observe the returns, which server as his observations. Hence, we let the observation space be $\mathscr{O} \triangleq \{1, 2\}$, where $o = 1$ represents a "failure" observation (return of zero) and $o = 2$ represents a "success" observation (return of one).[29] Similarly, the action space is $\mathscr{A} \triangleq \{1, 2\}$, where $a = 1$ represents the "continuing" action (staying with the current task) and $a = 2$ represents the "switching" action (drawing a new θ from $\mu(\theta)$). We let the ambiguity set (i.e., the "cloud" of models) be indexed by $m \in \mathscr{M} \triangleq \{1, 2, \cdots, |\mathscr{M}|\}$, where $\mu = \mu_m$ under model m. Using the notation of the general framework introduced in Section 3, for this job matching model, the core state transition probabilities under action $a = 1$ is given by $p^1_{ij}(m) \triangleq Pr\{j|i, a = 1, m\} = \mathbb{1}_{\{i=j\}}$ for all $i, j \in \mathscr{S}$, and under action $a = 2$ by $p^2_{ij}(m) \triangleq Pr\{j|i, a = 2, m\} = \mu_m(\theta_j)$. The observation transition probabilities under action $a = 1$ are defined by $r^1_{j1}(m) \triangleq Pr\{o = 1|j, a = 1, m\} = 1 - \theta_j$ and $r^1_{j2}(m) \triangleq Pr\{o = 2|j, a = 1, m\} = \theta_j$, and under action $a = 2$ as $r^2_{j1}(m) \triangleq Pr\{o = 1|j, a = 2, m\} = 1 - \mathbb{E}_{\mu_m}[\theta]$ and $r^2_{j2}(m) \triangleq Pr\{o = 2|j, a = 2, m\} = \mathbb{E}_{\mu_m}[\theta]$, where $\mathbb{E}_{\mu_m}[\theta] \triangleq \sum_{i \in \mathscr{S}} \theta_i \, \mu_m(\theta_i)$. Finally, we let

[26] This serially independent type of return is a common assumption in the literature (see, e.g., p. 311 of Stokey et al., 1989) that we also follow for simplicity. However, our APOMDP framework only requires Markovian dependencies, and can be used for modeling such extensions (e.g., when the worker's proficiency improves due to learning-by-doing).

[27] The majority of the literature considers the special case with $c = 0$ (see, e.g., Sections 10.10 and 10.11 of Stokey et al., 1989). Here, we allow for a general $c \in \mathbb{R}$. We also note that an extension to the case where the compensation depends on the match level is straightforward.

[28] Probability mass functions in $\{\mu_m : m = 1, 2, \cdots, |\mathscr{M}|\}$ need not be from the same family of distributions.

[29] The extension of our model to settings with more than two observations is straightforward. We use the binary observations here for simplicity and to match the literature (see, e.g., Sections 10.10 and 10.11 of Stokey et al., 1989).

the immediate reward under action $a \in \mathscr{A}$ be defined by $g^a \in \mathbb{R}^n$, where g^1 is a vector with ith element equal to θ_i, and g^2 is a vector with all elements equal to c.[30]

With these, we can write the underlying APOMDP Bellman equation using (8) as:

$$\begin{aligned} V_t(\boldsymbol{\pi},\alpha) &= \max_{a\in\mathscr{A}} \Big\{ \boldsymbol{\pi}' g^a + \beta\big[\alpha\, h_{t-1}(\boldsymbol{\pi},a,\underline{m}_{t-1}(\boldsymbol{\pi},a,\alpha),\alpha) \\ &\qquad + (1-\alpha)\, h_{t-1}(\boldsymbol{\pi},a,\overline{m}_{t-1}(\boldsymbol{\pi},a,\alpha),\alpha)\big]\Big\} \\ &= \max\Big\{ \mathbb{E}_{\boldsymbol{\pi}}[\theta] + \beta\big[\alpha\, h_{t-1}(\boldsymbol{\pi},1,\underline{m}_{t-1}(\boldsymbol{\pi},1,\alpha),\alpha) \\ &\qquad + (1-\alpha)\, h_{t-1}(\boldsymbol{\pi},1,\overline{m}_{t-1}(\boldsymbol{\pi},1,\alpha),\alpha)\big], \\ &\qquad c + \beta\big[\alpha\, h_{t-1}(\boldsymbol{\pi},2,\underline{m}_{t-1}(\boldsymbol{\pi},2,\alpha),\alpha) \\ &\qquad + (1-\alpha)\, h_{t-1}(\boldsymbol{\pi},2,\overline{m}_{t-1}(\boldsymbol{\pi},2,\alpha),\alpha)\big]\Big\}, \end{aligned} \tag{16}$$

and solve it along with the terminal condition $V_0(\boldsymbol{\pi},\alpha)=0$, which assumes (without loss of generality) that no return is collected at the end of the horizon. In (16), $\mathbb{E}_{\boldsymbol{\pi}}[\theta] \triangleq \sum_{i\in\mathscr{S}} \pi_i\,\theta_i$, and the function $V_t(\boldsymbol{\pi},\alpha)$ represents the optimal overall return of the worker (also referred to as the DM hereafter) when there are t periods to go, his belief is $\boldsymbol{\pi} \in \Pi_{\mathscr{S}}$, and his pessimism factor is α. Furthermore, similar to (7), in (16) we have

$$h_{t-1}(\boldsymbol{\pi},a,m,\alpha) \triangleq \sum_{o\in\mathscr{O}} Pr\{o|\boldsymbol{\pi},a,m\} V_{t-1}(T(\boldsymbol{\pi},a,o,m),\alpha), \tag{17}$$

where

$$Pr\{o|\boldsymbol{\pi},a,m\} = \sum_{i\in\mathscr{S}}\sum_{j\in\mathscr{S}} \boldsymbol{\pi}_i\, p^a_{ij}(m) r^a_{jo}(m) = \begin{cases} 1-\mathbb{E}_{\boldsymbol{\pi}}[\theta] & : o=1, a=1, \\ \mathbb{E}_{\boldsymbol{\pi}}[\theta] & : o=2, a=1, \\ 1-\mathbb{E}_{\mu_m}[\theta] & : o=1, a=2, \\ \mathbb{E}_{\mu_m}[\theta] & : o=2, a=2, \end{cases}$$

and the belief updating operator $T(\boldsymbol{\pi},a,o,m)$ can be calculated based on (2):

$$\begin{aligned} T(\boldsymbol{\pi},a,o,m) &= \frac{\big(\boldsymbol{\pi}' P^a_m R^a_m(o)\big)'}{Pr\{o|\boldsymbol{\pi},a,m\}} \\ &= \begin{cases} \hat{\boldsymbol{\pi}} \triangleq \big(\frac{\pi_1(1-\theta_1)}{1-\mathbb{E}_{\boldsymbol{\pi}}[\theta]}, \frac{\pi_2(1-\theta_2)}{1-\mathbb{E}_{\boldsymbol{\pi}}[\theta]}, \cdots, \frac{\pi_n(1-\theta_n)}{1-\mathbb{E}_{\boldsymbol{\pi}}[\theta]}\big)' & : o=1, a=1, \\ \tilde{\boldsymbol{\pi}} \triangleq \big(\frac{\pi_1\theta_1}{\mathbb{E}_{\boldsymbol{\pi}}[\theta]}, \frac{\pi_2\theta_2}{\mathbb{E}_{\boldsymbol{\pi}}[\theta]}, \cdots, \frac{\pi_n\theta_n}{\mathbb{E}_{\boldsymbol{\pi}}[\theta]}\big)' & : o=2, a=1, \\ \boldsymbol{\pi}^0_m \triangleq \big(\mu_m(\theta_1), \mu_m(\theta_2), \cdots, \mu_m(\theta_n)\big)' & : o=1, a=2, \\ \boldsymbol{\pi}^0_m \triangleq \big(\mu_m(\theta_1), \mu_m(\theta_2), \cdots, \mu_m(\theta_n)\big)' & : o=2, a=2. \end{cases} \end{aligned} \tag{18}$$

Thus, based on (17) we have:

$$h_{t-1}(\boldsymbol{\pi},a,m,\alpha) = \begin{cases} \big(1-\mathbb{E}_{\boldsymbol{\pi}}[\theta]\big) V_{t-1}(\hat{\boldsymbol{\pi}},\alpha) + \mathbb{E}_{\boldsymbol{\pi}}[\theta]\, V_{t-1}(\tilde{\boldsymbol{\pi}},\alpha) & : a=1, \\ V_{t-1}(\boldsymbol{\pi}^0_m,\alpha) & : a=2. \end{cases} \tag{19}$$

Therefore, the APOMDP Bellman equation (16) has a simple form:

$$\begin{aligned} V_t(\boldsymbol{\pi},\alpha) = \max\Big\{ &\mathbb{E}_{\boldsymbol{\pi}}[\theta] + \beta\big[\big(1-\mathbb{E}_{\boldsymbol{\pi}}[\theta]\big) V_{t-1}(\hat{\boldsymbol{\pi}},\alpha) + \mathbb{E}_{\boldsymbol{\pi}}[\theta]\, V_{t-1}(\tilde{\boldsymbol{\pi}},\alpha)\big], \\ &c + \beta\big[\alpha \min_{m\in\mathscr{M}} V_{t-1}(\boldsymbol{\pi}^0_m,\alpha) + (1-\alpha) \max_{m\in\mathscr{M}} V_{t-1}(\boldsymbol{\pi}^0_m,\alpha)\big]\Big\}. \end{aligned} \tag{20}$$

[30] To avoid the trivial case where receiving compensation c is optimal in each period, we shall assume c is small enough. In particular, we assume c is not greater than θ_n.

Finally, letting $\underline{m}$ and $\overline{m}$ be the minimizer and the maximizer in the second line of (20), respectively, we can write the APOMDP Bellman equation as:

$$V_t(\boldsymbol{\pi},\alpha)=\max\Big\{\mathbb{E}_{\boldsymbol{\pi}}[\theta]+\beta\big[\big(1-\mathbb{E}_{\boldsymbol{\pi}}[\theta]\big)V_{t-1}(\hat{\boldsymbol{\pi}},\alpha)+\mathbb{E}_{\boldsymbol{\pi}}[\theta]\,V_{t-1}(\tilde{\boldsymbol{\pi}},\alpha)\big],$$
$$c+\beta\big[\alpha V_{t-1}(\boldsymbol{\pi}^0_{\underline{m}},\alpha)+(1-\alpha)V_{t-1}(\boldsymbol{\pi}^0_{\overline{m}},\alpha)\big]\Big\}. \tag{21}$$

Importantly, we note that the first line of (21) (which corresponds to $a=1$) is independent of the model m, and $\underline{m}$ in the second line of (21) (which corresponds to $a=2$) is independent of the DM's belief, $\boldsymbol{\pi}$. Thus, the ambiguity set $\mathscr{M}$ has a BIWC member (see Definition 1). Therefore, the majority of the structural results developed in the previous sections for a general APOMDP hold for this setting. In the next section, we discuss such results in more detail, and show how they can be used to characterize the DM's value function and optimal policy.

6.1.1. Structural results: job matching with model ambiguity

It follows from Lemma 1 that the Bellman operator in (21) is a contraction mapping with modulus β on the complete metric space $(\mathscr{B}^{\alpha}, d^{\alpha})$. Hence, if we consider the job matching problem in infinite-horizon by setting $T=\infty$, taking the limit as $t\to\infty$ in (21), and denoting the infinite-horizon value function by $V_\infty(\boldsymbol{\pi},\alpha)$, from Proposition 1 we immediately have that for all $\pi\in\Pi_{\mathscr{S}}$ and $\alpha\in\mathscr{I}$ the value function $V_\infty(\boldsymbol{\pi},\alpha)$ is unique and satisfies $\lim_{t\to\infty}V_t(\boldsymbol{\pi},\alpha)=V_\infty(\boldsymbol{\pi},\alpha)$.

Since the ambiguity set $\mathscr{M}$ has a BIWC member, we observe from Proposition 2 that for all $t\in\mathscr{T}$ (i.e., for every period), $V_t(\boldsymbol{\pi},\alpha)$ defined in (21) is both piecewise-linear and convex in π for all α, and therefore, can be simply written as $V_t(\boldsymbol{\pi},\alpha)=\max_{\psi\in\Psi_{t,\alpha}}\{\boldsymbol{\pi}'\psi\}$. Furthermore, we note that based on the functions $p^a_{ij}(m)$ and $r^a_{jo}(m)$ defined earlier for the underlying job matching problem, the set of core state and observation transition kernels (P_m and R_m, respectively) are TP_2 (see Definition 4) for all $m\in\mathscr{M}$. Using this fact along with Theorem 1 and Proposition 5, we can establish the following monotonicity result.

Proposition 8 *(Job matching: monotonicity). The value function of the job matching problem under model ambiguity has the following monotonicity properties:*

(i) $V_t(\boldsymbol{\pi},\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$ for all $t\in\mathscr{T}$ and $\alpha\in\mathscr{I}$, and $V_\infty(\boldsymbol{\pi},\alpha)\in\mathscr{F}_{\Pi_{\mathscr{S}}}$ for all $\alpha\in\mathscr{I}$.

(ii) For all $\boldsymbol{\pi}\in\Pi_{\mathscr{S}}$ and $t\in\mathscr{T}$, $V_t(\boldsymbol{\pi},\alpha)$ is non-increasing in α. Similarly, for all $\boldsymbol{\pi}\in\Pi_{\mathscr{S}}$, $V_\infty(\boldsymbol{\pi},\alpha)$ is non-increasing in α.

The first part of the above result states that the value function in the job matching problem is TP_2-nondecreasing (see Definition 5) in both finite-horizon and infinite-horizon settings. This means that, all else equal, a higher belief (in the TP_2 sense) yields a (weakly) higher overall return for the worker. The second part of the above result states that, in both finite-horizon and infinite-horizon settings, a higher pessimism factor results in a (weakly) lower overall return. Hence, all else equal, the overall return cannot be higher for a more pessimistic worker than a more optimistic one.

Next, we turn our attention to the properties of the optimal policy. To provide insights into the structure of the optimal policy, we need the following definition of the set of real-valued TP_2-nonincreasing functions, which relies on Definition 5.

Definition 8 *(Real-valued TP_2-nonincreasing functions).* The set of real-valued TP_2-nonincreasing functions induced by $\Pi_{\mathscr{S}}$, denoted by $\hat{\mathscr{F}}_{\Pi_{\mathscr{S}}}$, 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 $f \in \hat{\mathscr{F}}_{\Pi_{\mathscr{S}}}$ if $-f \in \mathscr{F}_{\Pi_{\mathscr{S}}}$.

We also need the following definition whose power is in yielding monotone optimal policies (see, e.g., Topkis, 1998).

Definition 9 *(TP_2-supermodularity and TP_2-submodularity).* When $\mathscr{A} = \{1, 2\}$, the DM's APOMDP utility function defined in (9) is said to be TP_2-supermodular, if $U_t(\pi, \alpha, 2) - U_t(\pi, \alpha, 1) \in \mathscr{F}_{\Pi_{\mathscr{S}}}$ ($\forall t \in \mathscr{T}$, $\forall \alpha \in \mathscr{I}$). Similarly, when $\mathscr{A} = \{1, 2\}$, the DM's APOMDP utility function is said to be TP_2-submodular, if $U_t(\pi, \alpha, 2) - U_t(\pi, \alpha, 1) \in \hat{\mathscr{F}}_{\Pi_{\mathscr{S}}}$ ($\forall t \in \mathscr{T}$, $\forall \alpha \in \mathscr{I}$).

We next establish that for job matching problem with model ambiguity, the DM's utility function is TP_2-submodular.

Lemma 8 *(Job matching: TP_2-submodular utility). For the job matching problem with model ambiguity, the DM's utility function (in both finite-horizon and infinite-horizon settings) is TP_2-submodular.*

Since the utility function is TP_2-submodular, we can show that the optimal policy $a_t^* : \Pi_{\mathscr{S}} \times \mathscr{I} \to \mathscr{A}$ is nonincreasing (in the TP_2 sense). Thus, if the worker's optimal action is to continue with the current task ($a = 1$) when his belief is $\boldsymbol{\pi}$, his optimal action cannot be switching to a new task ($a = 2$) when his belief is $\boldsymbol{\pi}^+$, so long as $\boldsymbol{\pi} \preceq_{TP_2} \boldsymbol{\pi}^+$ (all else equal).

Theorem 3 *(Job matching: monotone optimal policy). In the finite-horizon setting, $a_t^*(\boldsymbol{\pi}, \alpha) \in \hat{\mathscr{F}}_{\Pi_{\mathscr{S}}}$ ($\forall t \in \mathscr{T}$, $\forall \alpha \in \mathscr{I}$). Similarly, in the infinite-horizon setting $a_\infty^*(\boldsymbol{\pi}, \alpha) \in \hat{\mathscr{F}}_{\Pi_{\mathscr{S}}}$ ($\forall \alpha \in \mathscr{I}$).*

In addition, it follows from Proposition 4 that the region for which action $a = 2$ is optimal is a convex set.

Proposition 9 *(Job matching: convex policy region). In the finite-horizon setting, $\Pi_{t,2}^*(\alpha) \triangleq \{\boldsymbol{\pi} \in \Pi_{\mathscr{S}} : a_t^*(\boldsymbol{\pi}, \alpha) = 2\}$ is a convex set ($\forall t \in \mathscr{T}$, $\forall \alpha \in \mathscr{I}$). Similarly, in the infinite-horizon setting, $\Pi_{\infty,2}^*(\alpha) \triangleq \{\boldsymbol{\pi} \in \Pi_{\mathscr{S}} : a_\infty^*(\boldsymbol{\pi}, \alpha) = 2\}$ is a convex set ($\forall \alpha \in \mathscr{I}$).*

The above results significantly reduce the complexity of characterizing the optimal policy. For example, it follows that at period t, there is a threshold surface $\Upsilon_t(\alpha) \subset \Pi_{\mathscr{S}}$ which divides the information space $\Pi_{\mathscr{S}}$ into two regions: one in which action $a = 1$ is optimal ($\Pi_{t,1}^*(\alpha)$) and one which action $a = 2$ is optimal ($\Pi_{t,2}^*(\alpha)$). A schematic illustration of these regions are presented in Fig. 4. Since $a_t^*(\boldsymbol{\pi}, \alpha)$ is monotone (Theorem 3), these results mean that $\Upsilon_t(\alpha)$ can be viewed as a *control-limit* on the DM's belief $\boldsymbol{\pi}$ which fully prescribes his optimal action. Furthermore, since by Proposition 9 the policy region $\Pi_{t,2}^*(\alpha)$ is a convex set, it follows that the threshold surface $\Upsilon_t(\alpha)$ is convex and almost everywhere differentiable. In addition, we can show that both policy regions $\Pi_{t,1}^*(\alpha)$ and $\Pi_{t,2}^*(\alpha)$ are connected sets[31] (see the proof of Proposition 11). Due to these results, the threshold surface $\Upsilon_t(\alpha)$ can be characterized using an effective approximation technique (see Section 6.2.1 for more details).

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

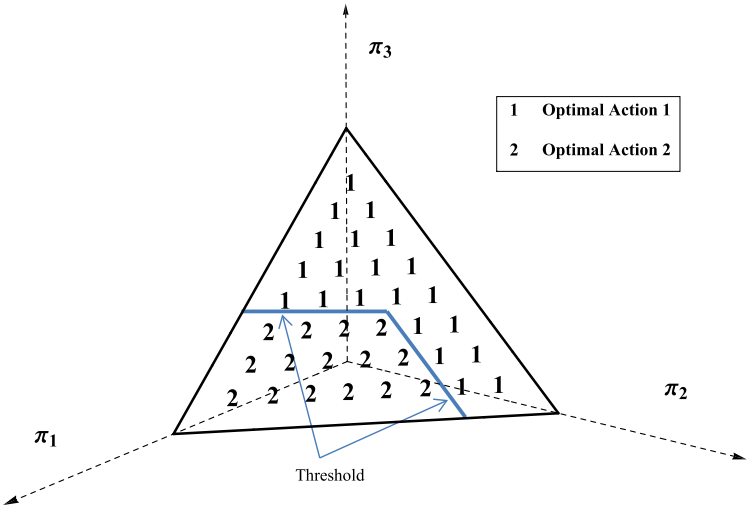


Fig. 4. A schematic representation of the optimal policy regions and the threshold in the job matching problem with model ambiguity ($n = 3$, $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$).

Finally, we note that the general results established in Section 5 can be used to provide a performance guarantee for a worker who is facing model ambiguity and follows the policy prescribed by solving (21). To this end, suppose one of the models in $\mathscr{M}$ is indeed the correct model. Not knowing this, and facing model ambiguity, the worker follows the policy prescribed by the APOMDP formulation (21). What is the maximum overall return (i.e., over the entire horizon) that he may lose compared to an imaginary version of himself who knows the correct model? We refer to this quantity as the worker's *price of ambiguity*, and provide an upper bound for it in the following proposition which is established using Theorem 2.

Proposition 10 *(Job matching: price of ambiguity). For any two models $m_1, m_2 \in \mathscr{M}$, define*

$$\gamma(\theta_j, m_1, m_2) \triangleq \max\Big\{ \big|\mu_{m_1}(\theta_j)(1 - \mathbb{E}_{\mu_{m_1}}[\theta]) - \mu_{m_2}(\theta_j)(1 - \mathbb{E}_{\mu_{m_2}}[\theta])\big|, \\ \big|\mu_{m_1}(\theta_j)\mathbb{E}_{\mu_{m_1}}[\theta] - \mu_{m_2}(\theta_j)\mathbb{E}_{\mu_{m_2}}[\theta]\big|\Big\}. \tag{22}$$

(i) The ambiguity set for the Job Matching Problem is ϵ^*-tight (see Definition 6), where $\epsilon^* \triangleq \max_{\theta_j, m_1, m_2} \gamma(\theta_j, m_1, m_2)$.

(ii) The worker's price of ambiguity over a horizon of T periods is no more than

$$\frac{3\xi^*\beta(3-\beta)(1-\beta^T)\theta_n}{(1-\beta)^3}, \tag{23}$$

where $\xi^* \triangleq \min\{2n\epsilon^*, \frac{3(1-\beta)}{3-\beta}\}$.

The result above suggests that the worker's price of ambiguity is relatively low: following the policy obtained from the APOMDP formulation enables the worker to make robust decisions. As (23) shows, the price of ambiguity is especially low for a worker who has a low ϵ^*, θ_n, and/or β.

6.2. Machine replacement problems with ambiguous deterioration

An important assumption in studies related to machine replacement problems is that machine deterioration probabilities (or that of any general asset under consideration) 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 optimal policy.

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

Definition 10 *(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^2_m and R^2_m are TP_2 for all $m \in \mathscr{M}$, (c) $p^1_{ij}(m) = \mathbb{1}_{\{j=s(m)\}}$ 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, we first show that the DM's APOMDP utility function for BAMMR APOMDPs is TP_2-Supermodularity (see Definition 9).

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.

Theorem 4 *(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 BIWC 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 $\alpha \in \mathscr{I}$. Furthermore, if $\mathscr{M}$ has a BIWC 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[32] on $\Pi_{\mathscr{S}}$. However, we can restrict our attention to TP_2 ordering on lines, which will resolve the issue.[33] To this end, let $e_i \in \Pi_{\mathscr{S}}$ represent a vector with a one as the ith element and zeros elsewhere, denote the convex hull of $e_1, e_2, \cdots, e_{n-1}$ by $\mathcal{C}$, and let $\mathscr{L}(\tilde{\boldsymbol{\pi}})$ be the line in $\Pi_{\mathscr{S}}$ that connects $\tilde{\boldsymbol{\pi}} \in \mathcal{C}$ to e_n:

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

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 $\pi, \hat{\pi} \in \mathscr{L}(\tilde{\boldsymbol{\pi}})$ (i.e., if π and $\hat{\pi}$ are on the same line connecting e_n to a point in $\tilde{\boldsymbol{\pi}} \in \mathcal{C}$). 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 5 is replaced with $\pi \preceq_{TP_2-L} \hat{\pi}$.

Theorem 4 enables us to state that for any BAMMR APOMDP (a) $a_t^*(\pi, \alpha) \in \mathscr{F}^L_{\Pi_{\mathscr{S}}}$, 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 such that $a_t^*(\pi, \alpha) = 1$ if π is in the first region and $a_t^*(\pi, \alpha) = 2$ otherwise,[34] even if the ambiguity set $\mathscr{M}$ does not have a BIWC member.

Proposition 11 *(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 BIWC 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.

6.2.1. Approximating the threshold

We now present a method to calculate the best linear approximation for the threshold $\Upsilon(\alpha)$ in any BAMMR APOMDP (a similar method can be used for the job-matching problem of

[32] 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.

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

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

Section 6.1 after some straightforward adjustments). 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:

$$a^{\hat{\Upsilon}(\alpha)}(\boldsymbol{\pi}, \alpha) = \begin{cases} 1 : \boldsymbol{\pi}' \hat{\Upsilon}(\alpha) \leq 1, \\ 2 : \boldsymbol{\pi}' \hat{\Upsilon}(\alpha) > 1. \end{cases} \tag{24}$$

The choice of 1 in the RHS of (24) 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 the same elements to an existing vector to make it positive.

Now, consider the optimization program

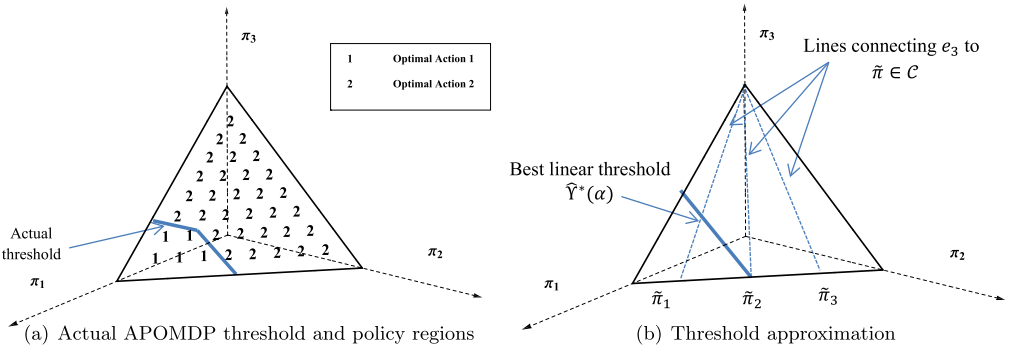
$$\begin{aligned} \hat{\Upsilon}^*(\alpha) = \arg \max_{\hat{\Upsilon}(\alpha) \in \mathbb{R}^n_+} \; & V_\infty^{\hat{\Upsilon}(\alpha)}(\cdot, \alpha) \\ \text{s.t.} \quad & ||\hat{\Upsilon}(\alpha)||_\infty = \hat{\Upsilon}_n(\alpha), \end{aligned} \tag{25}$$

where $V_\infty^{\hat{\Upsilon}(\alpha)}(\cdot, \alpha)$ denotes the infinite-horizon BAMMR APOMDP value function under the policy defined by (24). 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 (25), 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 12 *($TP_2 - L$ threshold).* $a^{\hat{\Upsilon}(\alpha)}(\boldsymbol{\pi}, \alpha) \in \mathscr{F}^L_{\Pi_\mathscr{S}}$ *if, and only if,* $||\hat{\Upsilon}(\alpha)||_\infty = \hat{\Upsilon}_n(\alpha)$.

Remark 5 *(Computing the approximate threshold).* Based on the result above, program (25) 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 (25), thereby characterizing an effective policy for the underlying BAMMR APOMDP. To this end, the objective function of program (12) 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 (25) can be easily transferred to an unconstrained one. For instance, one can use the change of variable $\hat{\Upsilon}_n(\alpha) = \left(\tilde{\Upsilon}_n(\alpha)\right)^2$ and $\hat{\Upsilon}_i(\alpha) = \left(\tilde{\Upsilon}_n(\alpha)\, \sin(\tilde{\Upsilon}_i(\alpha))\right)^2$. Then, the nonnegativity and the maximum-last-element requirements of program (25) 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



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

Fig. 5. Threshold approximation for a BAMMR APOMDP.

$s(1) = 1, s(2) = s(3) = 2$. 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 Fig. 5 (a). The best linear threshold obtained is depicted in Fig. 5 (b). As can be seen, the best linear threshold closely approximates the optimal threshold shown in Fig. 5 (a). Indeed, the approximation provides an effective control policy for the BAMMR APODMP problem. Fig. 5 (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 11 part (iii)).

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 $\boldsymbol{\pi} \in \Pi_{\mathscr{S}}$, player 1 chooses an action $a \in \mathscr{A}$ and receives a reward of $\boldsymbol{\pi}' g^a$. Simultaneously, a biased coin that has probability $\alpha \in \mathscr{I}$ of yielding head is tossed. The state of the system becomes $(\boldsymbol{\pi}, a, \omega)$, 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.[35] 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 $\boldsymbol{\pi} \in \Pi_{\mathscr{S}} \subset \bar{\Pi}_{\mathscr{S}}$, let player 1 action space be $\mathscr{A}$ and that of player 2 be an arbitrary singleton. For all $\boldsymbol{\pi} \in \Pi_{\mathscr{S}} \times \mathscr{A} \times \Omega \subset \bar{\Pi}_{\mathscr{S}}$, 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).

[35] 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).

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, 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 Smallwood and Sondik (1973), 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. For instance, the convexity of the value function in an APOMDP allows using similar computational methods to those already available for POMDPs (see, e.g., the discussion following Proposition 2) and the convexity of policy regions allows characterizing them via efficient optimization programs (see, e.g., Section 6.2.1).

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 representative numerical experiment. We show that, if hypothetically the DM

is allowed to optimize his pessimism level, he would not choose extreme values corresponding to maximax or maximin preferences. An important implication of this result is that it sheds light on the importance of deviating from a worst-case view of the world, which has widely appeared in the robust optimization literature.

We also discuss in detail the specific applications of APOMDPs and their structural results in (a) job matching, and (b) machine replacement problems. However, we note that APOMDPs can be used in a variety of other applications in economics and beyond, since in many such applications a decision-maker faces both hidden sates and model misspecification. One example that we briefly mention in some of our numerical experiments is that of a monetary authority who cannot fully observe the exact level of unemployment. He dynamically receives imperfect signals about the unemployment level and wants to accordingly set the inflation rate so as to control the unemployment level. For such an authority, the dynamics of unemployment rates under any given inflation rate is often ambiguous and cannot be defined via a single probabilistic model.

Other examples where APOMDPs can be applied include: *(1) Dynamic Principal-Agent Models:* Principal-agent models in dynamic settings have been extensively studied by economists since the 1970s (see, e.g., Holmstrom, 1979). POMDP-type models are used to address the underlying information asymmetry and/or moral hazard aspects in dynamic settings (see, e.g., Zhang and Zenios, 2008 and Saghafian and Chao, 2014). The APOMDP approach of this paper can extend such models by allowing the transition probabilities to be ambiguous instead of fully known. *(2) Stochastic Inventory Control:* Minimax and Bayesian solutions are vastly studied for inventory control problems perhaps starting from the early work of Scarf (1958) and Scarf (1959). Recently, a variety of papers have developed POMDPs to study inventory control for systems in which inventory is not fully observable (e.g., due to record inaccuracy). Unrealistically, however, the literature assumes that the demand distribution is fully known (see Saghafian and Tomlin, 2016 and the references therein for more discussions). The APOMDP approach proposed in this paper allows relaxing this assumption, and provides a method to develop inventory control strategies that do not rely on a particular demand distribution. *(3) Strategic Pricing and Revenue Management:* Strategic pricing and revenue management are widely studied in economics and related fields. Studies such as Aviv and Pazgal (2005) 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 and/or market dynamics. *(4) Medical Decision-Making:* POMDPs are widely used in the medical decision-making field. The patient's "health state" is typically not observable since medical tests are subject to false-positive and false-negative errors. 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. This typically results in estimation errors. In fact, in many medical decision-making applications, some actions (e.g., treatment options) are not commonly used by physicians in practice, and hence, there is only a very limited data (if any) regarding patient health transition probabilities under such actions. The APOMDP approach allows incorporating the resulted model misspecifications, and taking medical actions that are robust. *(5) Optimal Search:* Search for a hidden object is an important problem with various applications in economics, operations research, national security, and related fields. Existing models assume 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 hidden. APOMDPs provide a natural tool to find effective search policies without relying on a specific probabilistic model. *(6) Sequential Design of Experiments:* 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 thereby reduce the dependency of actions to unknown probability measures.

We leave it to future research to pursue the use of APOMDPs in various 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 above, we expect to see more results from future research in these directions.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jet.2018.08.006.

References

Ahn, D., Choi, S., Gale, D., Kariv, S., 2007. Estimating Ambiguity Aversion in a Portfolio Choice Experiment. Working Paper. UC Berkeley.

Hurwicz, L., 1997. An optimality criterion for decision making under ignorance. In: Arrow, K.J., Hurwicz, L. (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.

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

Bertsekas, D.P., Tsitsiklis, J.N., 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. In: 2nd Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 93–102.

Blackwell, D., 1951b. Comparison of experiments. Proc. Natl. Acad. Sci. 37, 826–831.

Blackwell, D., 1953. Equivalent comparison of experiments. Ann. Math. Stat. 24, 265–272.

Boloori, A., Saghafian, S., Chakkera, H.A., Cook, C.B., 2018. Data-Driven Management of Post-Transplant Medications: An APOMDP Approach. Working Paper. Harvard University.

Bren, A., Saghafian, S., 2016. Data-Driven Percentile Optimization for Multi-class Queueing Systems with Model Ambiguity: Theory and Application. Working Paper. Harvard University.

Cogly, T., Colacito, R., Hansen, L.P., Sargent, T.J., 2008. Robustness and U.S. monetary policy experimentation. J. Money Credit Bank. 40 (8), 1599–1623.

Cooper, R., Haltiwanger, J., 1993. The aggregate implications of machine replacement: theory and evidence. Am. Econ. Rev. 83 (3), 360–382.

de Farias, D.P., Van Roy, B., 2003. The linear programming approach to approximate dynamic programming. Oper. Res. 51 (6), 850–865.

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

Epstein, L.G., Schneider, M., 2003. Recursive multiple priors. J. Econ. Theory 113 (1), 1–31.

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

Ghirardato, P., Maccheroni, F., Marinacci, M., 2004. Differentiating ambiguity and ambiguity attitude. J. Econ. Theory 118, 133–173.

Ghirardato, P., Maccheroni, F., Marinacci, M., 2008. Revealed ambiguity and its consequences: updating. In: Abdellaoui, M., Hey, J.D. (Eds.), Advances in Decision Making under Risk and Uncertainty. Springer, pp. 3–18.

Gilboa, I., Schmeidler, D., 1989. Maxmin expected utility with non-unique prior. J. Math. Econ. 18 (2), 141–153.

Gupta, V., 2018. Near-Optimal Bayesian Ambiguity Sets for Distributionally Robust Optimization. Working Paper. University of Southern California.
Hansen, L.P., Sargent, T.J., 2001. Robust control and model uncertainty. Am. Econ. Rev. 91 (2), 60–66.
Hansen, L.P., Sargent, T.J., 2007. Recursive robust estimation and control without commitment. J. Econ. Theory 136, 1–27.
Hansen, L.P., Sargent, T.J., 2008. Robustness. Princeton University Press, Princeton, NJ.
Hansen, L.P., Sargent, T.J., 2012. Three types of ambiguity. J. Monet. Econ. 59, 422–445.
Heath, C., Tversky, A., 1991. Preference and belief: ambiguity and competence in choice under uncertainty. J. Risk Uncertain. 4 (1), 5–28.
Holmstrom, B., 1979. Moral hazard and observability. Bell J. Econ. 10, 74–91.
Hurwicz, L., 1951a. Optimality criteria for decision making under ignorance. In: Cowles Commission Discussion Paper. In: Statistics, vol. 370.
Hurwicz, L., 1951b. Some specification problems and applications to econometric models. Econometrica 19, 343–344.
Itoh, H., Nakamura, K., 2007. Partially observable Markov decision processes with imprecise parameters. Artif. Intell. 171, 453–490.
Iyengar, G.N., 2005. Robust dynamic programming. Math. Oper. Res. 30 (2), 257–280.
Jovanovic, B., 1979. Job matching and the theory of turnover. J. Polit. Econ. 87 (5), 972–990.
Jovanovic, B., 1982. Selection and the evolution of industry. Econometrica 50 (3), 649–670.
Jovanovic, B., Nyarko, Y., 1995. The transfer of human capital. J. Econ. Dyn. Control 19 (5), 1033–1064.
Jovanovic, B., Nyarko, Y., 1996. Learning by doing and the choice of technology. Econometrica 64 (6), 1299–1310.
Karlin, S., Rinott, Y., 1980. Classes of orderings of measures and related correlation inequalities: I. Multivariate totally positive distributions. J. Multivar. Anal. 10, 467–498.
Klibanoff, P., Marinacci, M., Mukerji, S., 2005. A smooth model of decision making under ambiguity. Econometrica 73 (6), 1849–1892.
Klibanoff, P., Marinacci, M., Mukerji, S., 2009. Recursive smooth ambiguity preferences. J. Econ. Theory 144, 930–976.
Krishnamurthy, V., Djonin, V., 2009. Optimal threshold policies for multivariate POMDPs in radar resource management. IEEE Trans. Signal Process. 57 (10), 3954–3969.
Krishnamurthy, V., Wahlberg, B., 2009. Partially observed Markov decision process multiarmed bandits – structural results. Math. Oper. Res. 34 (2), 287–302.
Lovejoy, W.S., 1987a. On the convexity of policy regions in partially observed systems. Oper. Res. 35 (4), 619–621.
Lovejoy, W.S., 1987b. Some monotonicity results for partially observed Markov decision processes. Oper. Res. 35 (5), 736–743.
Maccheroni, F., Marinacci, M., Rustichini, A., 2006. Dynamic variational preferences. J. Econ. Theory 128, 4–44.
Marinacci, M., 2002. Probabilistic sophistication and multiple priors. Econometrica 70 (2), 755–764.
Marschak, J., Miyasawa, K., 1968. Economic comparability of information systems. Int. Econ. Rev. 9, 137–174.
Monahan, G.E., 1982. A survey of partially observable Markov decision processes: theory, models, and algorithms. Manag. Sci. 28 (1), 1–16.
Nilim, A., El Ghaoui, L., 2005. Robust control of Markov decision processes with uncertain transition matrices. Oper. Res. 53 (5), 780–798.
Nowak, A.S., 1985. Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space. J. Optim. Theory Appl. 45 (4), 591–602.
Nowak, A.S., Szajowski, K., 1999. Nonzero-sum stochastic games. In: Bardi, M., Raghavan, T.E.S., Parthasarathy, T. (Eds.), Stochastic and Differential Games. In: Ann. Internat. Soc. Dynam. Games, vol. 4. Birkhäuser, Boston, pp. 297–342.
Osogami, T., 2015. Robust partially observable markov decision processes. In: Proceedings of the 32nd International Conf. on Machine Learning, vol. 37.
Riedel, F., 2009. Optimal stopping with multiple priors. Econometrica 77 (3), 857–908.
Rieder, U., 1991. Structural results for partially observed control models. Methods Models Oper. Res. 35, 473–490.
Ross, S., 1971. Quality control under Markovian deterioration. Manag. Sci. 17 (1), 587–596.
Saghafian, S., Chao, X., 2014. The impact of operational decisions on the optimal design of salesforce incentives. Nav. Res. Logist. 61 (4), 320–340.
Saghafian, S., Tomlin, B., 2016. The newsvendor under demand ambiguity: combining data with moment and tail information. Oper. Res. 64 (1), 167–185.
Scarf, H., 1958. A min–max solution of an inventory problem. Stud. Math. Theory Inventory Prod., 201–209.
Scarf, H., 1959. Bayes solution to the statistical inventory problem. Ann. Math. Stat. 30 (2), 490–508.
Shaked, M., Shanthikumar, J.G., 2007. Stochastic Orders. Springer, New York, NY.

Si, J., Barto, A.G., Powell, W.B., Wunsch, D., 2004. Handbook of Learning and Approximate Dynamic Programming. Wiley-IEEE Press.
Simon, R.S., 2007. The structure of non-zero-sum stochastic games. Adv. Appl. Math. 38 (1), 1–26.
Siniscalchi, M., 2011. Dynamic choice under ambiguity. Theor. Econ. 6, 379–421.
Smallwood, R., Sondik, E., 1973. The optimal control of partially observable Markov processes over a finite horizon. Oper. Res. 21, 1071–1088.
Smith, J.E., McCardle, K.F., 2002. Structural properties of stochastic dynamic programs. Oper. Res. 50 (5), 796–809.
Sondik, E., 1971. The Optimal Control of Partially Observable Markov Processes. Unpublished PhD dissertation. Stanford University.
Sondik, E.J., 1978. The optimal control of partially observable Markov processes over the infinite horizon: discounted costs. Oper. Res. 26 (2), 282–304.
Stein, C., 1951. Notes on a seminar on a theoretical statistics; comparison of experiments. Unpublished report.
Stokey, N.L., Lucas, R.E., Prescott, E.C., 1989. Recursive Methods in Economic Dynamics. Harvard University Press, Cambridge, MA.
Stoy, J., 2011. Statistical decisions under ambiguity. Theory Decis. 70 (2), 129–148.
Strzalecki, T., 2011. Axiomatic foundation of multiplier preferences. Econometrica 79 (8), 47–73.
Sulganik, E., 2003. On the structure of Blackwell's equivalence class of information systems. Math. Soc. Sci. 29 (3), 213–223.
Topkis, D.M., 1998. Supermodularity and Complementarity. Princeton University Press, Princeton, NJ.
Welch, L.R., 2003. Hidden Markov models and the Baum–Welch algorithm. IEEE Inf. Theory Soc. Newsl. 53 (4), 10–13.
White, C.C., 1978. Optimal inspection and repair of a production process subject to deterioration. J. Oper. Res. Soc. 29 (3), 235–243.
Whitt, W., 1980. Representation and approximation of noncooperative sequential games. SIAM J. Control Optim. 18 (1), 33–48.
Whitt, W., 1982. Multivariate monotone likelihood ratio and uniform conditional stochastic order. J. Appl. Probab. 19 (3), 695–701.
Wiesemann, W., Kuhn, D., Rustem, B., 2013. Robust Markov decision processes. Math. Oper. Res. 38 (1), 153–183.
Xu, H., Mannor, S., 2012. Distributionally robust Markov decision processes. Math. Oper. Res. 37 (2), 288–300.
Zhang, H., Zenios, S.A., 2008. A dynamic principal-agent model with hidden information: sequential optimality through truthful state revelation. Oper. Res. 56, 371–386.