Robust Partially Observable Markov Decision Processes

Mohammad Rasouli[1], Soroush Saghafian[2]

[1]Management Science and Engineering, Stanford University, Palo Alto, CA

[2]Harvard Kennedy School, Harvard University, Cambridge, MA

In a variety of applications, decisions needs to be made dynamically after receiving imperfect observations about the state of an underlying system. Partially Observable Markov Decision Processes (POMDPs) are widely used in such applications. To use a POMDP, however, a decision-maker must have access to reliable estimations of core state and observation transition probabilities under each possible state and action pair. This is often challenging mainly due to lack of ample data, especially when some actions are not taken frequently enough in practice. This significantly limits the application of POMDPs in real-world settings. In healthcare, for example, medical tests are typically subject to false-positive and false-negative errors, and hence, the decision-maker has imperfect information about the health state of a patient. Furthermore, since some treatment options have not been recommended or explored in the past, data cannot be used to reliably estimate all the required transition probabilities regarding the health state of the patient. We introduce an extension of POMDPs, termed Robust POMDPs (RPOMDPs), which allows dynamic decision-making when there is ambiguity regarding transition probabilities. This extension enables making robust decisions by reducing the reliance on a single probabilistic model of transitions, while still allowing for imperfect state observations. We develop dynamic programming equations for solving RPOMDPs, provide a sufficient statistic and an information state, discuss ways in which their computational complexity can be reduced, and connect them to stochastic zero-sum games with imperfect private monitoring.

Key words: Robust dynamic decision-making; Ambiguity; Imperfect state observation; Dynamic programming, Sufficient statistic; Information state; Stochastic zero-sum games.

History: Version: June 13, 2018

1. Introduction

Markov systems are systems in which, conditioned on the current state, the future state is independent of the past states, actions, and events. We study control of Markov systems where the decision maker ("DM" hereafter) (a) partially observes the state of the system (i.e., has imperfect state observations), (b) faces ambiguity about the transition distributions that define the dynamics of the system, and (c) uses maximin utility to make decisions that are robust to the ambiguity in transition distributions. We refer to such systems as *Robust Partially Observation Markov Decision Processes (RPOMDPs).*

RPOMDPs extend Partially Observable Markov Decisions Processes (POMDPs) by considering the important fact that defining transition probabilities via an exact probability distribution is often an impossible task. Hence, RPOMDPs can be viewed as an extension of POMDPs in which dynamic decisions need to be made without reliance on a single probabilistic model of transitions. This extension is important when we note that in contrast to the orthodox Bayesian paradigm, in most real-world applications, the DM does not have access to a single subjective or objective distribution for the underlying random variables. Instead, he must consider a set of possible distributions, and make decisions that perform well regardless of the actual distribution. This can be observed in a variety of applications including those in finance, healthcare, marketing, energy, and security systems, among others. For example, when analyzing investment opportunities, there is much ambiguity about market dynamics, and hence, defining future impact of current investments via a single probability distribution is often an impossible endeavor. When deciding about treatment options for a patient, evaluating the impact of a treatment strategy on the future health state of the patient is subject to ambiguity and cannot be perfectly quantified via a single probability distribution. Similarly, when choosing a marketing policy, there is much uncertainty about reception by potential clients, and thus, the impact of a given policy on the future well-being of a firm cannot be fully defined via a unique probability distribution.

Furthermore, in most such applications, the performance of a policy might be highly sensitive to the distribution of the underlying random variables. Therefore, it is essential to take into account the ambiguities surrendering such distributions, and empower the DM to make decisions that do not heavily rely on any particular probability distribution. RPOMDPs are particularly useful in this regard, because they allow for robust operations of the widely studied and used POMDPs.

In addition to POMDPs, RPOMPDs also provide an extension to Robust Markov Decision Processes (RMDPs), where a DM is not facing imperfect state observations, but has to deal with ambiguities surrounding transition probability distributions (see, e.g., Iyengar (2005), Nilim and El Ghaoui (2005), Wiesemann et al. (2013), and the references therein). Similar to RMDPs, the existence of ambiguity in RPOMDPs necessitates utilizing a robust control framework such as the maximin utility approach. However, for both RMDPs and RPOMDPs, the first common challenge is that the optimal decisions might not be dynamically consistent (see, e.g., Ellsberg's example in Ellsberg (1961)). Dynamic consistency is required for dynamic programming, and roughly speaking implies that the DM's preferences over alternatives do not change over time. Rectangularity is introduced in the literature as a sufficient condition for dynamic consistency under ambiguity (see, e.g., Maccheroni et al. (2006), Iyengar (2005), and Epstein and Schneider (2003)). This is because rectangularity ensures that, conditioned on the information available to the DM , the set of possible distributions of the future random variables does not depend on the previously chosen distributions (see Online Appendix B for more discussion on dynamic consistency and rectangularity). The second common challenge of RPOMDPs and RMDPs is that the information available to an adversarial agent ("nature" hereafter) for selecting the distribution of the random variables plays an important role. In particular, as we will show, if nature has perfect information about the DM's actions, then the DM cannot use a randomized policy to improve his utility over deterministic policies (see Cases 1 and 2 of Example 3 in Section 3.4). The third common challenge of RPOMDPs and RMDPs that we address in this paper is that their dynamic programming equations may require optimizing over all possible distributions of the future random variables.

Moreover, since RPOMDPs allow for both imperfect state observations (core advantage of POMDPs) and ambiguity (core advantage of RMDPs), they have a unique set of challenges that neither exists in POMPDs nor in RMDPs. The first challenge in this vein is with respect to modeling the DM's information about the previously selected distributions by nature. This information is important for characterizing the DM's optimal policy in RPOMDPs for two reasons (see Examples 1 and 2 in Section 3.3). First, this information can refine the DM's set of possible distributions over the current system states (see Example 2).[1] Second, the DM's information about nature's previous actions can change his belief about the set of possible future distributions (see Examples 1, Cases 1 and 2), unless the ambiguity set satisfies some restrictive conditions.[2] As a result of the above-mentioned two reasons, a correct modeling of the DM's information about the past distributions is highly important in RPOMDPs.

The second unique challenge of RPOMDPs is that, as we will show, the sufficient statistic for predicting the system's future behavior does not serve as an information state for its optimal control. This is in sharp contrast to both POMDPs and RMDPs. For example, a well-known result for POMDPs is that the DM's belief about the state of the system conditioned on his information history is both a sufficient statistic (for evaluating future behavior of the system) and an information state (for optimal control) (see, e.g., Bertsekas (1995)). Therefore, it is known that decision-making in a POMDP is equivalent to decision-making in an alternative system with perfect state observations, where the state of the new system is the DM's conditional belief over the states. Similarly, in RMDPs, under S-rectangularity[3] the state of the system (which is a sufficient statistic for evaluating the future state of the system) is also an information state for determining its optimal policy (see, e.g., Iyengar (2005) and Wiesemann et al. (2013)). Importantly, however, we find that this important result, which holds for both POMDPs and RMDPs, does not extend to RPOMDPs.

Specifically, we find that even under the assumption of S-rectangularity, the set of beliefs about the state in RPOMDPs (which is a sufficient statistic) cannot be used as an information state for optimal control (see Example 2 in Section 3.3). The reason behind this is that this sufficient statistic in RPOMDPs is not adequate for comparing the performance of different policies going

[1] This is not the case in RMDPs, where the state of the system can be uniquely determined by the DM without the need for knowing the selection of the past distributions by nature (see Example 1, Case 3).

[2] In general, this also holds in RMDPs although with some differences (see, e.g., Iyengar (2005)) .

[3] S-rectangularity means that the set of possible future distributions is independent of the past states, DM's actions, selection of the distribution of random variables, and values of the random variables conditioned on the current state of the system. For more information, see Online Appendix B as well as Witsenhausen (1966).

forward. Intuitively, this is due to the fact that the performance of each policy is the minimum total expected utility (sum of the past and future expected utilities) across different possible current state distributions and those of the future random variables. For two different policies, however, this minimum may be achieved by different distributions over the current state, which imply different expected utilities in the past. Hence, this information from the past should be added to the sufficient statistic to form an information state. Thus, we introduce one instance of an information state for RPOMDPs by adding the minimum expected utility for reaching each possible distribution of the system state to the sufficient statistic. Adding this "reward-to-reach" allows comparing the performance of policies starting from each of the possible current system state distributions (see Example 2 in Section 3.3).

The third unique challenge of RPOMDPs is their computational complexity, which is much higher than that of both RMDPs and POMDPs. The additional complexity in RPOMPDs compared to POMDPS and RMDPs has two sources: (a) a more complex policy space, and (b) a more complex state space. First, in contrast to POMDPs, deterministic policies are not necessarily optimal in RPOMDPs, and therefore, one needs to search the space of randomized policies which is much larger than that of deterministic policies. The second source of complexity—the state space complexity—is due to the ambiguity in the distribution of random variables that define the transitions of the system, which necessitates (a) optimizing over all possible distributions of the current states, (b) adding the reward-to-reach values to the state, and (c) optimizing over all possible transition probability distributions for dynamically updating the system's state.

To the best of our knowledge a complete model and study of RPOMDPs does not exist in the literature. We contribute by providing the first model and study in this regard. We do so by first introducing the notion of *ambiguous stochastic processes*: a sequence of random variables with an unknown joint distribution function that belongs to a set of possible joint distributions. We then use the notion of ambiguous stochastic processes to model *ambiguous input-output systems*. We next define and add a Markov property to introduce *ambiguous Markovian input-output systems*. Finally, we utilize a maximin expected utility approach in ambiguous Markovian input-output systems to rigorously define RPOMDPs.

After modeling RPOMDPs, we develop their dynamic programming equations by imposing a rectangularity assumption. However, we note that solving such equations is in general challenging, mainly because of the above-mentioned complexities in both the policy and the state space of RPOMDPs. Therefore, we next provide mechanisms to reduce such complexities. To reduce the complexity of the policy space, we show that although deterministic policies are not always optimal, if nature receives perfect observations of the DM's actions then there exists an optimal deterministic policy. This finding can be used to provide an approximation of the DM's optimal policy by reducing the policy space to the space of deterministic policies. To reduce the complexity of the state space, we consider the convex hull of an ambiguity set, and prove a *convex hull equivalence* theorem which enables us to significantly reduce the size of the ambiguity set. We show that, using this technique, the size of the ambiguity set can be reduced to the number of vertices of its convex hull. We also prove that, if the original ambiguity set is rectangular, then so is the set of its vertices.

Finally, we discuss the relation between RPOMDPs and stochastic zero-sum games in which players (a) both imperfectly and privately monitor each other's actions, and (b) have incomplete information of the state of the system. In particular, we propose a stochastic zero-sum game that has a value function equal to that of a corresponding RPOMDP. Thus, the dynamic programming equations we develop for RPOMDPs can also be used to solve the value function and maximin equilibrium strategies of such games. Alternatively, developing strategies for analyzing these equivalent games may enable finding efficient solutions for RPOMDPs.

The rest of the paper is organized as follows. In Section 2, we review the relevant studies. In Section 3, we develop a model for RPOMDPs, and in Section 4, we show how RPOMDPs can be analyzed via dynamic programming. In Section 5, we then propose a sufficient statistic and an information state for RPOMPDs. Next, we discuss various ways to reduce the computational complexity of RPOMDPs in Section 6. Finally, we present an equivalent game theoretic framework in Section 7, and briefly conclude in Section 8. All proofs can be found in the appendix

2. Literature

The axiomatic foundations of robust dynamic decision-making has been developed mainly in the economics literature, where different approaches for modeling ambiguity and for identifying a suitable decision criteria are studied. Besides maximin expected utility, other utility frameworks discussed in the economics literatures for decision-making under ambiguity include variational preferences, multiplier preferences, recursive smooth ambiguity, and α-maximin expected utility (see, e.g., Strzalecki (2011) for a related discussion and comparison of such approaches). Among these, we choose the maximin expected utility in this paper, where nature is assumed to be fully adversarial, and hence, the DM's utility of his decisions are determined based on worst-case outcomes.

The axiomatic foundations of robust dynamic decision-making has been studied by extending the static approaches. Dynamic minimax expected utility that we use in this paper is studied in Epstein and Schneider (2003), and can be considered as a generalization of the static maximin expected utility investigated by Gilboa and Schmeidler (1989). Hansen and Sargent (2005, 2007) study robust dynamic decision-making with multiplier preferences (i.e., entropy costs), where the decision maker has a best guess about the distribution of the random variables and assigns a cost to any distribution depending on how close it is to his best guess. Maccheroni et al. (2006) studies variational preferences, where similar to multiplier preferences the DM has a best guess distribution, but in contrast to Hansen and Sargent (2005, 2007), the costs assigned to distributions is not entropy-based. Klibanoff et al. (2009) uses the notion of smooth ambiguity preferences introduced in Klibanoff et al. (2005) (where the DM has a meta distribution over all possible distributions) and extends it to a dynamic setting. Another utilitarian framework in dealing with ambiguity is the α-maximin approach introduced by Marinacci (2002). Ghirardato et al. (2004) uses the α-maximin approach to shed light on separating the DM's ambiguity and his attitude toward it. Another relevant body of literature surrendering dynamic decision-making under ambiguity is concerned with the use of incoming data to dynamically overcome underlying ambiguities. For this stream of literature, we refers interested readers to Saghafian and Tomlin (2016), and the references therein.

The available studies on RMPDs are closely related to our work. This is mainly because RPOMDPs extend RMDPs by allowing for imperfect state observation. RMDPs have been studied in González-Trejo et al. (2002), Satia and Lave (1973), and later in Iyengar (2005) in parallel to Nilim and El Ghaoui (2005) under the assumption of state-action-rectangularity. This assumption requires the set of possible random variable distributions to be independent of the history condition on the state-action pairs. Wiesemann et al. (2013) study RMDPs under the weaker condition of S-rectangularity (state-rectangularity). This is the same notion of rectangularity that we also use in this paper. However, in contrast to this literature on RMDPs, we allow for partial observation of the system states (which covers RMDPs as a special case). This, in turn, brings new subtleties including the need for incorporating reward-to-reach values as part of the system state. The notion of reward-to-reach has appeared in the literature of risk sensitive MDPs or POMPDs (see, e.g., Bäuerle and Rieder (2013) and Bäauerle and Rieder (2017)). There, the DM's utility is a non-linear function of the accumulated utilities over time. Therefore, finding the optimal policy at each time requires the knowledge of the accumulated utility by that time (i.e., the reward-to-reach value).

RPOMDPs with maximin approach have also been studied under the condition that the set of state transition distributions consists of only degenerate distributions; that is, when each distribution refers to one specific transition almost surely (see, e.g., Bernhard (2000), Başar and Bernhard (2008), Bertsekas and Rhodes (1973), Rasouli et al. (2018), and Witsenhausen (1966)). Our model in this paper is much more general in that it allows for non-degenerate distributions. Similar to our work, Saghafian (2015) provides an extension of POMDPs, termed Ambiguous POMDPs (APOMDPs), by allowing for ambiguous transition probabilities. However, unlike the RPOMDP framework we study in this paper, the APOMDP approach proposed in Saghafian (2015) uses α-maximin preferences. Finally, a similar goal of extending POMDPs to enable decision-making under ambiguity is pursued in Itoh and Nakamura (2007), where it is assumed that the DM perfectly observes selections of distributions by nature in the past. In this paper, we consider both scenarios where the DM does or does not perfectly observe previous selections of distributions by nature. This allows us to shed light on the need to include reward-to-reach value as part of the information available to the DM.

3. The Model

Modeling RPOMDPs requires some preliminary definitions and results, which we present first. We then provide a model for RPOMDPs, and discuss the effects of the DM's and nature's information.[4]

We utilize the following notational convention throughout the paper. We denote random variables by capital letters (e.g., V), and their realizations by small letters (e.g., v). The dependency on time is denoted by subscript t (e.g., Z_t). $X_{t_1:t_2}$ is used to denote the vector $(X_{t_1}, X_{t_1+1}, ..., X_{t_2})$. Bold letters refer to sets (e.g., $\mathbf{\Gamma}$).

3.1. Preliminary Definitions

Definition 1 (**Ambiguous Random Variables**). A random variable Z with unknown probability distribution function $\gamma_Z \in \mathbf{\Gamma}_Z$, where $\mathbf{\Gamma}_Z$ is a given set of probability distribution functions, is said to be an *ambiguous random variable.* We refer to $\mathbf{\Gamma}_Z$ as the ambiguity set of Z, and may use the alternative notation $\mathbf{\Gamma}(Z)$ instead of $\mathbf{\Gamma}_Z$.[5]

Using measure theory, the above definition is equivalent to the following. Consider $\boldsymbol{P} = \{P^1, P^2, ..., P^n\}$ to be the set of probability functions measurable with respect to sample space $\mathbf{\Omega}$ and σ-algebra $\mathcal{F}$. Also, consider the random variable V as a measurable mapping from $(\mathbf{\Omega}, \mathcal{F})$ to $\mathbb{R}$ and Borel σ-algebra of $\mathbb{R}$, $\mathcal{B}_{\mathbb{R}}$. The possible distribution functions for random variable V corresponding to P^1, ..., P^n are γ^1, γ^2, ..., γ^n. Hence, $\mathbf{\Gamma}_Z = \{\gamma^1, \gamma^2, ..., \gamma^n\}$.

Definition 2 (**Uncertain Random Variables**). If $\mathbf{\Gamma}_Z$ is a singleton, then the ambiguous random variable Z is said to be an *uncertain random variable.*

The above definition builds a bridge between uncertainty and ambiguity (a.k.a. Knightian uncertainty) by presenting traditional random variables as a special case of their ambiguous counterparts. We can complete this bridge by defining ambiguous random processes, conditional ambiguous random variables, and independence of ambiguous random variables as follows.

Definition 3 (**Ambiguous Random Processes**). An *ambiguous random process* is a collection of ambiguous random variables $Z_{1:T} = (Z_1, Z_2, Z_3, ..., Z_T)$ with an unknown joint distribution

[4] In Online Appendix A, we discuss some real-world applications, and show how the appropriate modeling of the DM's and nature's information largely depends on the application.

[5] While the former notation provides a more concise presentation, the latter helps us to better represent conditional ambiguity.

function $\gamma_{Z_{1:T}} \in \mathbf{\Gamma}_{Z_{1:T}}$, where $\mathbf{\Gamma}_{Z_{1:T}}$ is the joint ambiguity set of the ambiguous random variables.[6]

Definition 4 (**Conditional Ambiguous Random Variables**). Consider event E in the space $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ (corresponding to event G in $(\mathbf{\Omega}, \mathcal{F})$). The ambiguous random variable Z conditioned on E has the conditional ambiguity set

$$\mathbf{\Gamma}(Z|E) = \{\gamma(Z|E) : \gamma \in \mathbf{\Gamma}_Z\}. \tag{1}$$

Similarly, for a subset of probability distribution functions $\mathbf{\Lambda} \subset \mathbf{\Gamma}_Z$, the conditional ambiguity set conditioned on E and $\mathbf{\Lambda}$ is[7]

$$\mathbf{\Gamma}(Z|E, \mathbf{\Lambda}) = \{\gamma(Z|E) : \gamma \in \mathbf{\Lambda}\}. \tag{2}$$

We next define the notions of weak and strong independence among ambiguous random variables. To this end, we first need to define the following set-function multiplication operator that operates on two sets of functions.

Definition 5 (**Set-Function Multiplication Operator**). The set-function multiplication operator $\underline{\times}$ operates on two sets of real-valued functions, set $\mathbf{A}$ of real-valued functions with domain $\mathbf{X}$ and set $\mathbf{B}$ of real-valued functions with domain $\mathbf{Y}$, and returns a new set of real-valued functions with domain $\mathbf{X} \times \mathbf{Y}$ consisting of the pairwise multiplication of the functions in $\mathbf{A}$ and $\mathbf{B}$.[8]

Definition 6 (**Weak Independence of Ambiguous Random Variables**). Consider ambiguous random variables $Z_{1:T} = (Z_1, Z_2, ..., Z_T)$ with joint ambiguity set $\mathbf{\Gamma}_{1:T}$. Consider a specific member of $\mathbf{\Gamma}_{1:T}$, $\gamma_{1:T}$, along with uncertain random variables $(\tilde{Z}_1, \tilde{Z}_2, ..., \tilde{Z}_T)$ that have the joint distribution function $\gamma_{1:T}$. If for all $\gamma_{1:T} \in \mathbf{\Gamma}_{1:T}$, the corresponding uncertain random variables $(\tilde{Z}_1, \tilde{Z}_2, ..., \tilde{Z}_T)$ are independent, then ambiguous random variables $Z_{1:T} = (Z_1, Z_2, ..., Z_T)$ are said to be *weakly independent.*

Definition 7 (**Strong Independence of Ambiguous Random Variables**). If for all $t = 1, ..., T-1$,

$$\mathbf{\Gamma}(Z_{1:T}) = \mathbf{\Gamma}(Z_{1:t}) \underline{\times} \mathbf{\Gamma}\big(Z_{t:T}\big), \tag{3}$$

then ambiguous random variables $Z_{1:T} = (Z_1, Z_2, ..., Z_T)$ are said to be *strongly independent.*

[6] For notational convenience, and whenever it is clear which random variables we are referring to (e.g. ambiguous random variables $Z_{1:T}$), we may use the simplified notations $\gamma_{1:T}$ and $\mathbf{\Gamma}_{1:T}$ in place of $\gamma_{Z_{1:t}}$ and $\mathbf{\Gamma}_{Z_{1:T}}$, respectively. Similarly, we may use γ_t and $\mathbf{\Gamma}_t$ to refer to γ_{Z_t} and $\mathbf{\Gamma}_{Z_t}$, respectively.

[7] Note that $(E, \mathbf{\Lambda})$ can be interpreted as an event in the space $\big(\mathbb{R} \times \mathbf{\Gamma}, \mathcal{B}_{\mathbb{R}} \otimes \sigma(\mathbf{\Gamma})\big)$, where $\sigma(\mathbf{\Gamma})$ is the sigma algebra generated by $\mathbf{\Gamma}$.

[8] As an example, if $\mathbf{A} = \{f_1\}$, and $\mathbf{B} = \{g_1, g_2\}$, where $f_2 : \mathbf{X} \to \mathbb{R}$ and $g_1, g_2 : \mathbf{Y} \to \mathbb{R}$, then $\mathbf{A} \underline{\times} \mathbf{B} = \{f_1 \times g_1, f_1 \times g_2\}$.

Theorem 1 **(Decomposition under Strong Independence)**. *If ambiguous random variables $Z_{1:T} = (Z_1, Z_2, ..., Z_T)$ with ambiguity set $\mathbf{\Gamma}_{1:T}$ are strongly independent, then $\mathbf{\Gamma}_{1:T} = \mathbf{\Gamma}_1 \underline{\times} \mathbf{\Gamma}_2 \underline{\times} ... \underline{\times} \mathbf{\Gamma}_T$.*

Theorem 2 **(Strong vs. Weak Independence)**. *If ambiguous random variables $Z_{1:T} = (Z_1, Z_2, ..., Z_T)$ are strongly independent, then they are also weakly independent. The converse is not necessarily true.*

3.2. Modeling Robust POMDPs

To model Robust POMDPs (RPOMDPs), we consider an input-output system with (a) a finite time horizon $\boldsymbol{T} = \{1, 2, ..., T\}$, (b) state $s_t \in \boldsymbol{S}_t$ at time t, where $\boldsymbol{S}_t$ is a finite set, and (c) an ambiguous random process $(Z_1, Z_2, ..., Z_T)$ with a joint ambiguity set $\mathbf{\Gamma}_{1:T}$ that determine state and observation dynamics.

The sequence of events in this input-output system is depicted in Figure 1 and is as follows.[9] Initially, at time $t = 1$, the state of the system is $s_1 \in \boldsymbol{S}_1$ and known to both the DM and nature. At each time $t = 1, ..., T-1$, first the DM selects action $a_t \in \boldsymbol{A}_t$. Nature then receives her observation $y_t^n \in \boldsymbol{Y}_t^n$:

$$y_t^n = g_t^n(s_t, a_t, z_{t-1}, x_t), \tag{4}$$

which is a function of the state of the system at time t (i.e., s_t), the DM's action at time t (i.e., a_t), and the realized value of the random variable Z_{t-1} (i.e., z_{t-1}).[10] This observation is imperfect and distorted by nature's observation noise X_t, which is an uncertain random variable with a realized value x_t based on a known distribution γ_{X_t}.

Next, nature selects $\gamma_t \in \mathbf{\Gamma}_t$ which is the distribution of the random variable $Z_t = (V_t, W_t)$, where W_t is an ambiguous random variable affecting the DM's observation (see (5)) and V_t is an ambiguous random variable affecting the system's state transitions (see (6)). The value of Z_t (denoted by $z_t = (v_t, w_t)$) is then realized based on the nature's selected distribution $\gamma_t \in \mathbf{\Gamma}_t$. Then, the DM receives his observation y_t:

$$y_t = g_t(s_t, a_t, \gamma_t, w_t), \tag{5}$$

which is a function of the system state s_t, DM's action a_t, nature's selected distribution γ_t, and the realized value of the DM's observation noise w_t.

Next, the system state transitions to s_{t+1} as a function of s_t, a_t and v_t:

$$s_{t+1} = f_t(s_t, a_t, v_t). \tag{6}$$

[9] Note that the realization of stage utilities is also depicted in this figure, which we will discuss later in this section.

[10] For the special case of y_1^n, assume z_0 is a parameter with its value known to both the DM and nature.

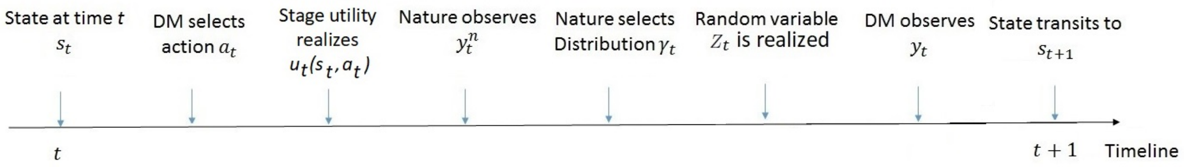


Figure 1 Sequence of Events in an RPOMDP

The collection of primary random variables in this model is $\mathbf{R}=(X_1,Z_1,X_2,Z_2,...,X_{T-1},Z_{T-1})$, which is an ambiguous random process with a joint ambiguity set denoted by $\mathbf{\Gamma}_{\boldsymbol{R}}$.

At time $t\in\boldsymbol{T}$, we denote the DM's and nature's policy for choosing their actions by σ_t and ψ_t, respectively. The information available to the DM at time t is denoted by

$$h_t=\{z_0,s_1,a_{1:t-1},y_{1:t-1}\}, \tag{7}$$

with a corresponding information space denoted by $\boldsymbol{H}_t$. Note that

$$h_{t+1}=(h_t,a_t,y_t). \tag{8}$$

The DM's policy at time t is a function from his information space $\boldsymbol{H}_t$ to the space of his (potentially) randomized actions denoted by $\Delta(\boldsymbol{A}_t)$:

$$\sigma_t:\boldsymbol{H}_t\rightarrow\Delta(\boldsymbol{A}_t), \quad t=1,2,...,T, \tag{9}$$

where $\Delta(\boldsymbol{A}_t)$ is the simplex generated by all the convex combinations of his set of deterministic actions $\boldsymbol{A}_t$. The set of all functions defined in (9) forms the set of DM's admissible policies at time t, which we denote by $\mathbf{\Sigma}_t$. Thus, $\mathbf{\Sigma}_{1:T-1}$ denotes DM's policy over the entire horizon.

Similarly, we denote the information available to nature at time t by

$$h_t^n=\{z_0,s_1,\gamma_{1:t-1},y_{1:t}^n\}, \tag{10}$$

with a corresponding information space denoted by $\boldsymbol{H}_t^n$. Nature's policy at time t, ψ_t, is then:

$$\psi_t:\boldsymbol{H}_t^n\rightarrow\mathbf{\Gamma}(Z_t|h_t^n), \quad t=1,2,...,T. \tag{11}$$

where $\mathbf{\Gamma}(Z_t|h_t^n)$ is the set of possible distributions for random variable Z_t given h_t^n. The set of all functions defined in (11) forms the set of nature's admissible policies at time t, which we denote by $\mathbf{\Psi}_t$. Finally, $\mathbf{\Psi}_{1:T-1}$ denotes the nature's policy space over the entire horizon.

We refer to the system modeled above via Equations (4)-(11) as an *ambiguous input-output system*. We next define a Markov property for this system using the equivalent definition for traditional uncertain input-output systems (see, e.g., Bertsekas (1995)).

Definition 8 (**Markovian Ambiguous Input-Output Systems**). Consider the ambiguous input-output system modeled above via (4)-(11) with an ambiguity set $\boldsymbol{\Gamma}_{\boldsymbol{R}}$. For every $\hat{\gamma}_{\boldsymbol{R}} \in \boldsymbol{\Gamma}_{\boldsymbol{R}}$, construct a counterpart uncertain input-output system (a system without any ambiguity) in which uncertain random variables $\hat{Z}_{1:T}$ have a joint distribution $\hat{\gamma}_{1:T}$, the DM's observations are given by $\hat{y}_t = g_t(s_t, a_t, \hat{\gamma}_t, \hat{w}_t)$, and the state transition is defined by $s_{t+1} = f_t(s_t, a_t, \hat{v}_t)$. If for all $\hat{\gamma}_{\boldsymbol{R}} \in \boldsymbol{\Gamma}_{\boldsymbol{R}}$, this constructed uncertain input-output system is Markovian (i.e., the future system states are independent of the past conditioned on the current system state (Bertsekas 1995)), then the ambiguous input-output system is said to be Markovian.

Finally, using the preliminaries above, we can define an RPOMDP. Consider the DM's stage utility function to be $u_t(s_t, a_t)$ at time $t = 1, 2, ..., T-1$ and $u_T(s_T)$ at time T. The total utility from $t = 1$ to $t = T$ is the discounted sum of the stage utilities with a discount factor $\beta \in [0, 1]$. If the DM was facing only uncertainty and not ambiguity (i.e., if $|\boldsymbol{\Gamma}_{1:T}| = 1$), then his expected total utility generated by policy $\sigma_{1:T-1}$ would be

$$E^{\sigma_{1:T-1}}_{Z_{1:T-1}}\Big[\sum_{t=1,\dots,T-1} \beta^t u_t(s_t, a_t) + \beta^T u_T(s_T)\Big]. \tag{12}$$

Since in an RPOMDP the DM is facing ambiguity, (12) can take different values depending on the distribution of $Z_{1:T-1}$ (i.e., $\gamma_{1:T-1}$). We assume that the DM is conservative, and wants to maximize the minimum of such values. That is, his objective is to maximize his minimum total expected utility.[11]

Definition 9 (**RPOMDPs**). An RPOMDP is a Markovian ambiguous input-output system defined via (4)-(11) in which the DM's objective is to find a policy that maximizes his minimum total expected utility:

$$\sigma^*_{1:T-1} = \arg \max_{\sigma_{1:T-1} \in \boldsymbol{\Sigma}_{1:T-1}} \min_{\psi_{1:T-1} \in \boldsymbol{\Psi}_{1:T-1}} E^{\sigma_{1:T-1}, \psi_{1:T-1}}_{A_{1:T-1}, Z_{1:T-1}, X_{1:T-1}} \Big\{ \sum_{t=1,2,\dots,T-1} \beta^t u_t(s_t, a_t) + \beta^T u_T(s_T) \Big\} \tag{13}$$

$$s.t. \quad y^n_t = g^n_t(s_t, a_t, z_{t-1}, x_t) \quad 1 \leq \forall t \leq T-1, \tag{14}$$

$$y_t = g_t(s_t, a_t, \gamma_t, w_t) \quad 1 \leq \forall t \leq T-1, \tag{15}$$

$$s_{t+1} = f_t(s_t, a_t, v_t) \quad 1 \leq \forall t \leq T-1. \tag{16}$$

We define a rectangular RPOMDP as a special case of an RPOMDP that satisfies rectangularity.

Definition 10 (**Rectangular RPOMDPs**). An RPOMDP with a rectangular ambiguity set is said to be a rectangular RPOMDP.

[11] As noted earlier, various alternative approaches to maximin for dealing with ambiguity are proposed in the literature. We refer interested readers to Maccheroni et al. (2006) and the references therein for a discussion on such approaches. Here, we adopt the maximin approach for its simplicity and theoretical attractiveness.

A special case of a rectangular RPOMDP is an S-rectangular RPOMDP which we define below.

Definition 11 (**S-Rectangular RPOMDPs**). A rectangular RPOMDP with an S-rectangular ambiguity set is said to be a S-rectangular RPOMDP.

In Sections 4 and 5, we use the notions of rectangular and S-rectangular RPOMDPs defined above to (a) develop dynamic programming equations, and (b) provide a sufficient statistic and an information state.

3.3. The Impact of the DM's Information

In (5), the DM's observation is not only a function of the state s_t, the DM's action a_{t-1}, and the realized value of the random variable W_t, but also the nature's action γ_t (i.e., the selected distribution of the random variable Z_t). In this section, we present two examples (Examples 1 and 2) to show the impact and importance of the information available to the DM regarding the nature's actions. These examples highlight two important effects. First, the information about the selected distributions by nature in previous periods can refine future ambiguity sets, and hence, change the DM's optimal policy. We examine this effect by considering an RMDP (as a special case of RPOMDPs defined in Definition 9) under perfect and imperfect knowledge about nature's actions via Cases 1 and 2 of Example 1, respectively. However, as we show via Case 3 of Example 1, this effect does not exist if the ambiguity is rectangular. Second, the information about the nature's selection of distributions in past can change the DM's current belief about the system state. We show this second effect via Example 2 by considering a rectangular RPOMDP both with and without the DM's ability to observe nature's actions.

Since the DM's information about the nature's action (i.e., selected γ_t) plays an important role, prior to presenting Examples 1 and 2, we categorize the DM based on this information into two categories: perfect and imperfect.

Definition 12 (**Perfect and Imperfect DM**). If a DM can perfectly observe the selection of the distribution γ_t by nature at all $t \in \boldsymbol{T}$, then he is said to be a perfect DM. A DM who is not perfect is called an imperfect DM.

For future use, we also define a special case of an imperfect DM: absolutely imperfect DM.

Definition 13 (**Absolutely Imperfect DM**). An imperfect DM is said to be absolutely imperfect, if his observations are independent of the nature's selection γ_t at all $t \in \boldsymbol{T}$, i.e., if

$$y_t = g_t(s_t, a_t, w_t). \tag{17}$$

We denote the probability distribution of the system state at time t by π_t, and the ambiguity set surrendering it by $\boldsymbol{\Pi}_t$. Note that a perfect DM has enough information (knows $\gamma_{1:t-1}$ exactly) to form a unique distribution over the system states as his belief at time t.[12]. An imperfect DM, however, cannot do so, because he does not have access to $\gamma_{1:t-1}$. Therefore, his belief about the system state at time t is not unique: it belongs to a set of possible system state distributions. That is, while for a perfect DM we have $|\mathbf{\Pi}_t| = 1$, for an imperfect DM we might have $|\mathbf{\Pi}_t| > 1$. Moreover, when the DM is imperfect, for each member of $\mathbf{\Pi}_t$ there exists a corresponding $\gamma_{1:t} \in \mathbf{\Gamma}_{1:t}$ that might have been selected by nature.

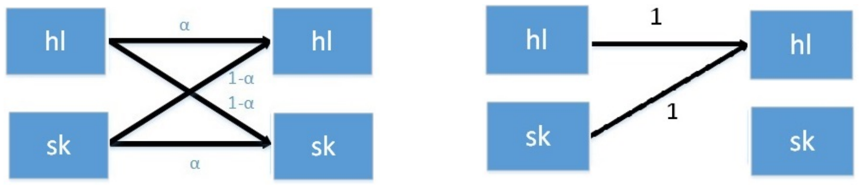
With these, we can make use of the following two stylized examples to show the important effect of the DM's information. The first example, Example 1, is presented in three separate cases.

Example 1 (**DM's Information - First Effect**). Consider treatment of a patient by a physician (DM) modeled as a stylized two-stage process. At each stage, the patient can be in one of two states labeled as healthy (hl) and sick (sk), and the physician can take one of the two actions labeled as treat (tr) and not treat (nt). Based on the physician's action, the patient transitions to a new state probabilistically as shown in Figure 2 with probabilities determined by parameter α, where the value of α is $\alpha = \alpha_1$ at time $t = 1$, and $\alpha = \alpha_2$ at time $t = 2$. The physician is ambiguous about the transition probabilities, and the value of α_t ($t = 1, 2$) is chosen by nature from an ambiguity set. Furthermore, at each stage, s/he performs a medical test prior to taking an action, which yields an observation about the patient's health state. This observation depends on the patient's health state and the selected distribution by nature. The physician's stage utilities as a function of the patient's health state and his/her action are given in Table 1. The physician's total utility is the sum of his/her stage utilities, where for simplicity it is assumed that $\beta = 1$.

Example 1, Case 1 (RMDP with Perfect DM). In this case, we assume that the physician's observations of the patient's health state are perfect. That is, the medical test has negligible false-positive and false-negative errors. Furthermore, we assume that the physician is a perfect DM: s/he fully observes the value of α_t ($t = 1, 2$) selected by nature. These two assumptions imply that, in this case, the general RPOMDP of Example 1 is simplified to an RMDP with a perfect DM. To further simplify this case, suppose that the physician's ambiguity set for α_t is $(\alpha_1, \alpha_2) \in \{(0.9, 0.9), (0.5, 0.5)\}$, and therefore, the physician knows that $\alpha_1 = \alpha_2$.

To show the effect of the DM's information, consider the physician's problem at the beginning of the second stage ($t = 2$). Suppose s/he has observed that $\alpha_1 = 0.5$ and $s_2 = sk$. Since s/he has observed $\alpha_1 = 0.5$, s/he can infer that $\alpha_2 = 0.5$. Thus, if s/he does not treat, his expected future utility is $u_2(sk, nt) + [\alpha_2 u_3(sk) + (1 - \alpha_2) u_3(hl)] = (-10) + [0.5 \times (0) + 0.5 \times (-10)] = -15$, but if s/he treats, her/his expected future utility is $u_2(sk, tr) + [0 \times u_3(sk) + 1 \times u_3(hl)] = (-16) + [0 \times$

[12] The reader should note that this is a distribution over system states, and the exact system state is still hidden to the DM.



(a) State transition probabilities when the physician does not treat the patient.

(b) State transition probabilities when the physician treats the patient.

Figure 2 State transitions as a function of the DM's (physician's) action in Example 1

Table 1 The DM's stage utilities in Example 1 at times $t = 1, 2, 3$, $u_1(s_1, a_1)$, $u_2(s_2, a_2)$ and $u_3(s_3)$.

	sk, tr	sk, nt	hl, tr	hl, nt	sk	hl
u_1	-20	-14	-50	0	-	-
u_2	-16	-10	-4.5	0	-	-
u_3	-	-	-	-	-10	0

Table 2 The DM's expected future utility from time $t = 2$ in Example 1, Case 2.

Action	$\alpha_2 = 0.9$	$\alpha_2 = 0.5$
nt	$(-10) + [0.9 \times (-10) + 0.1 \times (0)] = -19$	$(-10) + [.5 \times (-10) + 0.5 \times (0)] = -15$
tr	$(-16) + [0 \times (-10) + 1 \times (0)] = -16$	$(-16) + [0 \times (-10) + 1 \times (0)] = -16$

$(-10) + 1 \times (0)] = -16$. Therefore, the physician's best action at the second stage when $s_2 = sk$ is not to treat, nt. However, we show in the next case that this is not the optimal action, if the physician cannot observe the nature's selection.

Example 1, Case 2 (RMDP with Absolutely Imperfect DM). Consider a similar setting to Case 1, but assume that the physician is an absolutely imperfect DM: s/he cannot observe α_1, α_2, or any information about them. In this setting, the physician remains ambiguous about α_2 when trying to make a decision at the second stage. The expected future utility of each action for each possible value of α_2 at the second stage is shown in Table 2. Clearly the physician's best action in this case is to treat, which is in contrast to his/her best action in Case 1 above. This observation highlights the first way in which the information available to the DM with respect to the nature's action can make a difference in the optimal policy: it can affect the DM's ambiguity about future distributions. However, as we will show in the next case (Case 3), this effect disappears whenever the ambiguity is rectangular, so long as we impose the simplifying assumptions discussed in Case 1 (i.e., if we limit the setting to an RMDP). Importantly, however, via Example 2 we will show that in RPOMDPs (which are more general than RMDPs), even when the ambiguity set is rectangular, the information available to the DM with respect to the nature's action can alter the optimal policy. This occurs because the information available to the DM in RPOMDPs can change his expected reward-to-reach to each possible conditional system state distribution.

Example 1, Case 3 (RMDP with Rectangular Ambiguity). Consider the setup in Case 2 with the difference that $\alpha_1 \in \{0.5, 0.9\}$ and $\alpha_2 = \{0.5, 0.9\}$, and that α_1 and α_2 are not necessarily equal (i.e., $(\alpha_1, \alpha_2) \in \{(0.5, 0.5), (0.5, 0.9), (0.9, 0.9), (0.9, 0.5)\}$). Note that here the knowledge about α_1 does not limit the set of possible values for α_2. Specifically, this ambiguity set is strongly independent (see Definition 7), which induces a rectangularity property (see Online Appendix B). It is easy to see that, in this case, the expected future utility of different actions at the second stage ($t = 2$) is not affected by the information about α_1. Moreover, if the DM's observation at the second stage, y_2, is also independent of α_2, then the DM's ambiguity set about α_2 at the second stage is $\alpha_2 = \{0.5, 0.9\}$. Therefore, the DM's utilities are the same as those in Table 2, which means that the DM's best action is to treat.

Example 2 (**DM's Information - Second Effect**). We present this example as a continuation of the setup in Example 1 with the following changes. Assume that the medical test used has false-positive and/or false-negative errors, and hence, the physician has imperfect information about the health state of the patient. His belief at $t = 1$ about the health state of the patient is denoted by $\pi_1 = (\pi_1(hl), \pi_1(sk))$. Due to the underlying ambiguities in state and/or observation transition probabilities, the physician is ambiguous about π_1, and cannot fully specify it. For simplicity, suppose that $\pi_1 \in \big\{(1,0), (0,1)\big\}$, the physician decides to treat at $t = 1$, and his/her observation at $t = 1$ is independent of the patient's health state. Let us assume that the transition distributions from $t = 1$ to $t = 2$ following treatment at $t = 1$ are as shown in Figure 3 (note that this is different from the setup in Example 1). However, assume that the transition distribution from $t = 2$ to $t = 3$ is the same as the one in Case 3 of Example 1. With this simple setting in mind, we can show that only considering the future utility is not sufficient for determining the optimal policy at $t = 2$. To see this, let us first determine the physician's belief about the patient's health state at $t = 2$. If π_1 was $(1, 0)$, then $\pi_2 = (1, 0)$, but if $\pi_1 = (0, 1)$, then $\pi_2 = (0.5, 0.5)$. Therefore, the physician's belief about s_2 is $\pi_2 \in \mathbf{\Pi}_2 = \{(1, 0), (0.5, 0.5)\}$. Next observe that the sum of the physician's utilities at times $t = 2$ and $t = 3$ when his/her second stage decision is not to treat is as presented in Table 3.[13] The maximin utility from not treating the patient is -10. When treating the patient, the utilities are as given in Table 4. Clearly, the maximin utility of treating is -10.25. Therefore, the best action by considering only the future utilities is not to treat the patient at time $t = 2$. However, we can see that this is not the physician's best action considering all the information available to him/her. To observe this, note that when deciding about an action at time $t = 2$, the physician's objective is to maximize the minimum total utility from $t = 1$ to $t = 3$,

[13] For example, from this table, the utility with $\pi_2 = (0.5, 0.5)$ under $\alpha_2 = 0.9$ is $0.5 \times \big((0) + [0.9 \times (0) + 0.1 \times (-10)]\big) + 0.5 \times \big((-10) + [0.1 \times (0) + 0.9 \times (-10)]\big) = -10$, and under $\alpha_2 = 0.5$ is $0.5 \times \big((0) + [0.5 \times (0) + 0.5 \times (-10)]\big) + 0.5 \times \big((-10) + [0.5 \times (-10) + 0.5 \times (0)]\big) = -10$.

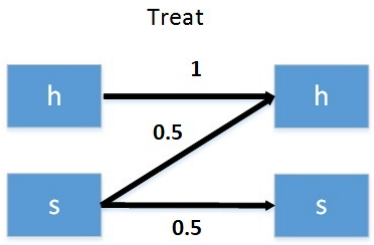


Figure 3 Transition distributions for Example 2 at time $t=1$ when the physician treats.

Table 3 Sum of the future utilities at time $t=2$ for not treating in Example 2.

	$\alpha_2=0.9$	$\alpha_2=0.5$
$\pi_2=(1,0)$	$(0)+[0.9\times(0)+0.1\times(-10)]=-1$	$(0)+[0.5\times(0)+0.5\times(-10)]=-5$
$\pi_2=(0.5,0.5)$	-10	-10

Table 4 Sum of the future utilities at time $t=2$ for treating in Example 2.

	$\alpha_2=0.9$	$\alpha_2=0.5$
$\pi_2=(1,0)$	$(-4.5)+[1\times(0)+0\times(-16)]=-4.5$	$-4.5+1\times(0)+0\times(-16)=-4.5$
$\pi_2=(0.5,0.5)$	$0.5\times(-4.5)+0.5\times(-16)=-10.25$	$0.5\times(-4.5)+0.5\times(-16)=-10.25$

Table 5 Sum of the past and future utilities at time $t=2$ for not treating in Example 2.

	$\alpha_2=0.9$	$\alpha_2=0.5$
$\pi_2=(1,0)$	-50-1=-51	-50-5=-55
$\pi_2=(0.5,0.5)$	-20-10=-30	-20-10=-30

Table 6 Sum of the past and future utilities at time $t=2$ for treating in Example 2.

	$\alpha_2=0.9$	$\alpha_2=0.5$
$\pi_2=(1,0)$	-50-4.5=-54.5	-50-4.5=-54.5
$\pi_2=(0.5,0.5)$	-20-10.25=-30.25	-20-10.25=-30.25

not only the utility from $t=2$ to $t=3$. In fact, the physician knows that if $\pi_2=(1,0)$, then π_1 should have been $(1,0)$, and therefore, s/he has already achieved a utility of $u_1(hl,tr)=-50$. But if $\pi_2=(0.5,0.5)$, then π_1 should have been $(0,1)$, and therefore, his/her gained utility from $t=1$ is $u_1(sk,tr)=-20$. Considering these past utilities, the total past and future utilities under different actions at $t=2$ are given in Tables 5 and 6. From these tables, it is clear that not treating at $t=2$ has a maximin utility of -55, while treating at $t=2$ has a maximin utility of -54.5. Therefore, the physician's optimal action is to treat at time $t=2$. This simple example illustrates that, unlike RMPDs, in RPOMDPs even when the ambiguity set is rectangular, the DM's optimal action may depend on the information available to the DM with respect to the nature's action. This is due to different expected utilities in the past or reward-to-reach values that might be achieved, and is the second way we discussed earlier in which information about the nature's actions can alter the DM's optimal policy.

Remark 1 (**Learning**). In many dynamic decision-making scenarios, learning may occur when the DM faces ambiguity. Through learning, a DM can reduce his ambiguity by making use of his observations. In our setting, the ambiguity is with respect to the underlying distributions that define the dynamics of the system. If the DM can infer that some distributions have not been used by nature in the past, then he might be able to use this information to learn about future distributions. For example, if a specific $\hat{\gamma}_{1:t-1} \in \mathbf{\Gamma}_{1:t-1}$ is not feasible conditioned on the DM's information h_t at time t, then any $\gamma_{1:T-1} \in \mathbf{\Gamma}_{1:T-1}$ which is in the form of $\gamma_{1:T-1} = \{\hat{\gamma}_{1:t-1}, \gamma_{t:T-1}\}$ cannot be in $\mathbf{\Gamma}_{1:T-1}(h_t) = \mathbf{\Gamma}(Z_{t:T-1}|h_t)$. Thus, this information helps the DM to refine the set $\mathbf{\Gamma}_{t:T-1}(h_t) = \mathbf{\Gamma}(Z_{t:T-1}|h_t)$. More generally, the DM might be able to infer feasibility of the past distributions $\hat{\gamma}_{t-1} \in \mathbf{\Gamma}_{1:t-1}$ directly or indirectly. If the DM is not absolutely imperfect (see Definition 13), he can learn about the past distributions directly from his observations (Example 1, Case 1). An absolutely imperfect DM cannot learn in the same direct way, because his observations are not a direct function of the previously selected distributions by nature, $\gamma_{1:t-1}$ (Example 1, Case 2). However, an absolutely imperfect DM can learn indirectly by considering the dependency of his observation function to random variables $w_t, t \in \{1, 2, ..., T\}$. If he observes an event that has a zero probability under a certain distribution $\hat{\gamma}_{1:t-1}$, then he can infer that $\hat{\gamma}_{1:t-1}$ has not been selected in the past.

3.4. The Impact of Nature's Information

In addition to the DM's information, the information available to nature can also change the DM's optimal policy. We show this via Example 3 below. Before doing so, however, we first categorize nature based on her information into two categories: perfect adversarial and imperfect adversarial. A special case of the imperfect adversarial nature is absolutely imperfect adversarial nature, which is also defined below.

Definition 14 (**Perfect and Imperfect Adversarial Nature**). If nature perfectly observes the DM's past actions (i.e., $\{a_{1:t-1}\} \in h_t^n$), then she is said to be perfect adversarial. When nature is not perfect adversarial, she is said to be imperfect adversarial. A special case of imperfect adversarial nature is absolutely imperfect adversarial nature, which refers to the scenario where nature's information history is limited to her own actions (i.e., $h_t^n = \{\gamma_{1:t-1}\}$). An absolutely imperfect nature does not observe any of the DM's actions either directly or indirectly through realizations of random variables.

Using the following example, we next highlight the impact that the nature's information can have on the DM's policy.

Example 3 (**Effect of Nature's Information**). Consider an RPOMDP with two stages. The system is initially at state s_1 right before the DM takes his action. The DM takes an action followed by an action from nature. Consequently, the system transitions to either s_2^1 or s_2^2. The DM's set of actions is $A = \{a_1^1, a_1^2\}$. There is a terminal utility of $u_2(s_2^1) = -100$ and $u_2(s_2^2) = 0$. The transition distributions can be based on either of the following two cases depending on nature's selection: $P(s_2^1|a_1^1, s_1) = 1$ and $P(s_2^1|a_1^2, s_1) = 0$, or $P(s_2^1|a_1^1, s_1) = 0$ and $P(s_2^1|a_1^2, s_1) = 1$.

Example 3, Case 1 (Perfect Adversarial Nature). Clearly if nature observes the DM's action before choosing the transition distribution, she can always force the system to move to s_2^1 by choosing $P(s_2^1|a_1^1, s_1) = 1$ when a_1^1 is chosen by the DM, and $P(s_2^1|a_1^1, s_1) = 0$ when a_1^2 is chosen by the DM. This means that nature can enforce the lower utility to the DM. Note that in this case the DM cannot improve his utility by using a *randomized* policy, because even if he uses a randomized policy, the realization of his action is observed by nature. In Theorem 8, we show that this result holds for any general RPOMDP with perfect adversarial nature. Therefore, when the nature is perfect adversarial, one can always search for an optimal policy for the DM within the set of deterministic policies.

Example 3, Case 2 (Imperfect Adversarial Nature). However, if nature does not observe the DM's actions, the nature cannot enforce the lower utility to the DM. In this case, the DM can improve his performance by randomizing between his actions. Similarly, the nature's policy that minimizes the DM's utility is also to randomize between the two distributions. That is, both the DM and nature can benefit from a randomized policy.

Cases 1 and 2 of Example 3 show how the information available to the nature can affect the optimal policy of both the DM and nature. We take advantage of the insights gained via this and other representative examples presented earlier, and provide a set of general results in the following sections.

4. Dynamic Programming

In this section, we develop dynamic programming (a.k.a. Bellman) equations for RPOMDPs. Dynamic programming equations can be used for two purposes: (1) evaluating the utility of a certain policy (policy evaluation), and (2) searching among all policies in the policy space for finding the optimal policy. To be able to write and use the dynamic programming equations, we assume that the rectangularity condition holds which, in turn, serves as a sufficient condition for dynamic consistency (see Online Appendix B).

To formalize the dynamic programming equations, we introduce three values for the DM under a given policy $\sigma_{1:T-1}$: total utility, reward-to-reach, and reward-to-go. Total utility of a given policy evaluated at time t denoted by $U_t^{\sigma_{t:T-1}}$ is the expected discounted utility accumulated under that policy during the entire horizon (i.e., from 1 to T). Reward-to-reach of a given policy denoted by $R_t^{\sigma_{t:T-1}}$ is the expected discounted utility accumulated under that policy up to but excluding time

t (i.e., from 1 to $t-1$). Finally, reward-to-go of a certain policy denoted by $V_t^{\sigma_{t:T-1}}$ (a.k.a the value function), is the expected discounted reward that will be accumulated under that policy during the rest of the horizon (i.e., from t to T). Clearly, at any time t, the total utility under a given policy can be calculated as the sum of the reward-to-reach and reward-to-go, and when comparing policies, it is important to consider this sum. However, as we show below, comparing two policies can be done by only considering their reward-to-go values when the DM is perfect, and hence, dynamic programming equations can be simplified in this case. As we will see, however, dynamic programming equations are more complex for an RPOMDP with an imperfect DM.

For a perfect DM, the total utility under a policy $\sigma_{1:T}$ at time t (conditioned on DM's information history h_t) can be calculated as:

$$U_t^{\sigma_{t:T-1}}(h_t) = \min_{\psi_{t:T-1}\in\boldsymbol{\Psi}_{t:T-1}(h_t)} E^{\sigma_{t:T-1},\psi_{t:T-1}}_{A_{t:T-1},Z_{1:T-1},X_{1:T-1}}\left\{\sum_{\tau=1,2,\ldots,T-1}\beta^\tau u_\tau(s_\tau,a_\tau)+\beta^T u_T(s_T)\middle|h_t\right\} \quad (18)$$
$$s.t. \qquad (14)-(16),$$

where the minimization is over all possible nature's policies in the future (conditioned on h_t) denoted by $\psi_{t:T-1}\in\boldsymbol{\Psi}_{t:T-1}(h_t)$. It is important to note that the expectation in (18) is independent of both the DM's policy in the past, $\sigma_{1:t-1}$, and nature's policy in the past, $\psi_{1:t-1}$. This is because it is conditioned on information history h_t defined in (7), and the realizations of both the DM's actions, $a_{1:t-1}$, and nature's actions, $\gamma_{1:t-1}$, exist in h_t. Hence, regardless of what the policy has been, the DM knows the realization of the past actions for both himself and nature. Thus, as the left hand side of (18) indicates, the DM's total utility at time t given h_t only depends on his policy from time t onward (and not his policy in the past). Also, since the realization of the DM's actions in the past $a_{1:t-1}$ is available in h_t, the expectation in (18) is only over $A_{t:T-1}$. This is not true for $Z_{1:t-1}$ and $X_{1:t-1}$, because although the perfect DM knows their distributions at time t, he does not know their realized values.

We now show how $U_t^{\sigma_{t:T-1}}(h_t)$ formulated in (18) can be written as the sum of the reward-to-reach and reward-to-go values. A perfect DM knows the selected distribution of the random variables in the past, $\gamma_{1:t-1}$. Therefore, he can calculate a unique reward-to-reach value at any time t based on his information, $R_t(h_t)$, as:[14]

$$R_t(h_t) = E_{Z_{1:t-1},X_{1:t-1}}\left\{\sum_{\tau=1,2,\ldots,t-1}\beta^\tau u_\tau(s_\tau,a_\tau)\middle|h_t\right\} \quad (19)$$

$$s.t. \quad y^n_\tau = g^n_\tau(s_\tau,a_\tau,z_{\tau-1},x_\tau) \qquad 1\le\forall\tau\le t-1, \quad (20)$$

$$y_\tau = g_\tau(s_\tau,a_\tau,\gamma_\tau,w_\tau) \qquad 1\le\forall\tau\le t-1, \quad (21)$$

[14] The reader should note that, based on the discussion following (18), $R_t(h_t)$ is independent of the DM's and nature's policies, since it is conditioned on h_t.

$$s_{\tau+1} = f_\tau(s_\tau, a_\tau, v_\tau) \qquad 1 \le \forall \tau \le t-1. \tag{22}$$

To calculate the total utility, the DM's reward-to-go value under policy $\sigma_{t:T-1}$, $V_t^{\sigma_{t:T-1}}(h_t)$, should be added to $R_t(h_t)$. This reward-to-go value, $V_t^{\sigma_{t:T-1}}(h_t)$, can be calculated via minimization over all $\psi_{t:T-1} \in \mathbf{\Psi}_{t:T-1}(h_t)$:

$$\begin{aligned} V_t^{\sigma_{t:T-1}}(h_t) = & \min_{\psi_{t:T-1} \in \mathbf{\Psi}_{t:T-1}(h_t)} E^{\sigma_{t:T-1}, \psi_{t:T-1}}_{A_{t:T-1}, Z_{t:T-1}, X_{t:T-1}, S_t} \Bigg\{ \sum_{\tau = t, t+1, \dots, T-1} \beta^{\tau - t} u_\tau(s_\tau, a_\tau) \\ & + \beta^{T-t} u_T(s_T) \Bigg| h_t \Bigg\}, \qquad (23) \\ s.t. \qquad & (14) - (16). \end{aligned}$$

Note that the state of the system, S_t, is not perfectly known to the DM. Therefore, the expectation in (23) would not be well-defined without considering the dependency on S_t. However, since the perfect DM knows $\gamma_{1:t-1}$, he can use Bayesian updating to form a unique conditional distribution for the state of the system, $\pi_t(h_t)$.[15]

Once the reward-to-reach and reward-to-go values are calculated, we can compute the total utility of a perfect DM as

$$U_t^{\sigma_{t:T-1}}(h_t) = R_t(h_t) + \beta^t V_t^{\sigma_{t:T-1}}(h_t), \tag{24}$$

which holds under any policy $\sigma_{t:T-1}$. From (24), and since $R_t(h_t)$ is independent of the perfect DM's policy, we observe the following important insight. When the DM is perfect, finding the DM's optimal policy from time t onward requires optimization only over reward-to-go values:

$$\sigma^*_{t:T-1} = \arg \max_{\sigma_{t:T-1} \in \mathbf{\Sigma}_{t:T-1}} V_t^{\sigma_{t:T-1}}(h_t). \tag{25}$$

We now consider the case of imperfect DM and seek to find out whether (25) holds in this case as well. The challenge with an imperfect DM is that he does not know $\gamma_{1:t-1}$ at time t. Therefore, an imperfect DM can neither form a unique distribution over the system states (e.g., by using Bayesian updating) nor a unique reward-to-reach values. This results in the following two subtleties both of which necessitate optimizing over all possible past distributions, $\gamma_{1:t-1} \in \mathbf{\Gamma}_{1:t-1}$.[16] First, in the case of imperfect DM, (24) does not hold, and needs to be modified as:

$$U_t^{\sigma_{t:T-1}}(h_t) = \min_{\gamma_{1:t-1} \in \mathbf{\Gamma}_{1:t-1}} U_t^{\sigma_{t:T-1}}(h_t, \gamma_{1:t-1}) \tag{26}$$

[15] Although $V_t^{\sigma_{t:T-1}}$ is the sum of the utilities from t to $T-1$, to form $\pi_t(h_t)$, the DM needs to consider the system dynamics from 1 to $t-1$. This results in complexity in the dynamic programming equations. In next sections, we will propose an information state that can be used to overcome this complexity.

[16] Note that conditioning on $\gamma_{1:t-1}$ essentially means forming a perfect DM.

$$= \min_{\gamma_{1:t-1} \in \mathbf{\Gamma}_{1:t-1}} R_t(h_t, \gamma_{1:t-1}) + \beta^t V_t^{\sigma_{t:T-1}}(h_t, \gamma_{1:t-1}), \tag{27}$$

where, for any given $\gamma_{1:t-1}$, the reward-to-reach and reward-to-go values in (27) can be calculated similar to their counterparts for the case of a perfect DM.[17] Second, from (27) we observe that to find the optimal policy from time t onward, $\sigma^*_{t:T-1}$, one should consider the total utility. This is because for comparing two policies, in contrast to a perfect DM, the imperfect DM cannot simply compare their reward-to-go values as they may be due to two different $\gamma_{1:t-1}$, and hence, two different reward-to-reach values (see Example 2). As a result, when the DM is imperfect, (25) needs to be modified as:

$$\begin{aligned} \sigma^*_{t:T-1} &= \arg \max_{\sigma_{t:T-1} \in \mathbf{\Sigma}_{t:T-1}} U_t^{\sigma_{t:T-1}}(h_t) \\ &= \arg \max_{\sigma_{t:T-1} \in \mathbf{\Sigma}_{t:T-1}} \min_{\gamma_{1:t-1} \in \mathbf{\Gamma}_{1:t-1}} U_t^{\sigma_{t:T-1}}(h_t, \gamma_{1:t-1}). \end{aligned} \tag{28}$$

In what follows, we consider the above-mentioned subtleties, and develop the underlying dynamic programming formulations separately for RPOMDPs with a perfect DM and with an imperfect DM.

4.1. Dynamic Programming for RPOMDPs with a Perfect DM

We start by considering the case where the DM is perfect. As noted earlier, in this case the reward-to-reach values can be excluded from the calculation of the optimal policy from time t onward. Let $V_t^*(h_t)$ denote the maximum value of all possible DM policies from time t to $T-1$:

$$\begin{aligned} V_t^*(h_t) = &\max_{\sigma_{t:T-1} \in \mathbf{\Sigma}_{t:T-1}} \min_{\psi_{t:T-1} \in \mathbf{\Psi}_{t:T-1}(h_t)} E^{\sigma_{t:T-1}, \psi_{t:T-1}}_{A_{t:T-1}, Z_{t:T-1}, X_{t:T-1}, S_t} \Bigg\{ \sum_{\tau = t, t+1, \ldots, T-1} \beta^{\tau - t} u_\tau(s_\tau, a_\tau) \\ &+ \beta^{T-t} u_T(s_T) \Bigg| h_t \Bigg\} \\ &s.t. \qquad (14) - (16), \end{aligned} \tag{29}$$

where the DM's distribution over S_t is $\pi_t(h_t)$. The DM's policy $\sigma^*_{t:T-1}$ corresponding to $V_t^*(h_t)$ is then the optimal policy from time t onward.

THEOREM 3 **(Dynamic Programming for RPOMDPs with a Perfect DM)**. *For a rectangular RPOMDP with a perfect DM, functions $V_t^*, t \in \boldsymbol{T}$ defined by (29) satisfy the following Bellman equations. At time $t = T$*

$$\begin{aligned} V_T^*(h_T) &= E_{S_T}\{u_T(s_T) \big| h_T\} \\ s.t. \quad &(14) - (16), \end{aligned} \tag{30}$$

[17] Following Remark 1, in the above discussion the optimization over $\gamma_{1:t-1} \in \mathbf{\Gamma}_{1:t-1}$ can be replaced by $\gamma_{1:t-1} \in \mathbf{\Gamma}_{1:t-1}(h_t)$ where $\mathbf{\Gamma}_{1:t-1}(h_t) = \mathbf{\Gamma}(Z_{t:t-1}|h_t)$. Depending on how much information the DM receives about $\gamma_{1:t-1}$, $\mathbf{\Gamma}_{1:t-1}(h_t)$ can be a much smaller set than $\mathbf{\Gamma}_{1:t-1}$.

and at time $t < T$,

$$V_t^*(h_t) = \max_{\sigma_t \in \boldsymbol{\Sigma}_t} \min_{\psi_t \in \boldsymbol{\Psi}_t(h_t)} E_{A_t, Z_t, X_t, S_t}^{\sigma_t, \psi_t} \left\{ u_t(s_t, a_t) + \beta V_{t+1}^*(h_{t+1}) \middle| h_t \right\} \tag{31}$$

$$s.t. \quad y_\tau^n = g_\tau^n(s_\tau, a_\tau, z_{\tau-1}, x_\tau) \qquad 1 \le \forall \tau \le t, \tag{32}$$

$$y_\tau = g_\tau(s_\tau, a_\tau, \gamma_\tau, w_\tau) \qquad 1 \le \forall \tau \le t, \tag{33}$$

$$s_{\tau+1} = f_\tau(s_\tau, a_\tau, v_\tau) \qquad 1 \le \forall \tau \le t. \tag{34}$$

The advantage of using the result provided by Theorem 3 compared to (29) is that one only needs to optimize over σ_t and ψ_t instead of $\sigma_{t:T-1}$ and $\psi_{t:T-1}$. Furthermore, the system dynamics constraints are only considered from time 1 to t (instead of 1 to T).[18]

4.2. Dynamic Programming for RPOMDPs with an Imperfect DM

We now develop the dynamic programming formulation for RPOMDPs when the DM is imperfect. As noted earlier, an imperfect DM cannot form a unique distribution over the system state, because he does not know $\gamma_{1:t-1}$. This necessitates (a) optimizing over all possible distributions of the random variables in the past (i.e., all possible values of $\gamma_{1:t-1}$), and (b) utilizing both reward-to-reach and reward-to-go values.

The following theorem demonstrates that, similar to the case with a perfect DM, the set of functions $V_t^*, t \in \boldsymbol{T}$ for RPOMDPs with an imperfect DM can be calculated via dynamic programming. Since the imperfect DM does not know $\gamma_{1:t-1}$, to present this theorem, we need to consider the selected distributions by nature from time 1 to $t-1$ that result in the maximin total utility at time t (denoted by $\gamma_{1:t-1}^*$), as well as the DM's and nature's optimal policy from t onward corresponding to it (denoted by $\sigma_{t:T-1}^*$ and $\psi_{t:T-1}^*$, respectively):

$$\begin{aligned}(\sigma_{t:T-1}^*, \psi_{t:T-1}^*, \gamma_{1:t-1}^*) = \arg &\max_{\sigma_{t:T-1} \in \boldsymbol{\Sigma}_{t:T-1}} \min_{\psi_{t:T-1} \in \boldsymbol{\Psi}_{t:T-1}(h_t)} \min_{\gamma_{1:t-1} \in \boldsymbol{\Gamma}_{1:t-1}} \\ E_{A_{t:T-1}, Z_{1:T-1}, X_{1:T-1}}^{\sigma_{t:T-1}, \psi_{t:T-1}, \gamma_{1:t-1}} &\left\{ \sum_{\tau=1,2,\dots,T-1} \beta^\tau u_\tau(s_\tau, a_\tau) + \beta^T u_T(s_T) \middle| h_t \right\} \\ &s.t. \quad (14)-(16).\end{aligned} \tag{35}$$

This enables us to formulate the imperfect DM's reward-to-go value (value function) under $(\sigma_{t:T-1}^*, \psi_{t:T-1}^*, \gamma_{1:t-1}^*)$ as

$$\begin{aligned}V_t^*(h_t) = E_{A_{t:T-1}, Z_{t:T-1}, X_{t:T-1}, S_t}^{\sigma_{t:T-1}^*, \psi_{t:T-1}^*, \gamma_{1:t-1}^*} &\left\{ \sum_{\tau=t,t+1,\dots,T-1} \beta^{\tau-t} u_\tau(s_\tau, a_\tau) + \beta^{T-t} u_T(s_T) \middle| h_t \right\} \\ s.t. \quad (14)-(16).&\end{aligned} \tag{36}$$

[18] These dynamic programming equations are complex because they require considering the system dynamics from 1 to $t-1$. We will introduce sufficient statistic and information state in Section 5 to reduce this complexity.

Theorem 4 **(Dynamic Programming for RPOMDPs with an Imperfect DM)**. *For an RPOMDP with imperfect DM, functions $V_t^*(t \in \boldsymbol{T})$ defined by (36) satisfy the following Bellman equations. At time $t = T$,*

$$V_T^*(h_T) = \min_{\gamma_{1:T} \in \boldsymbol{\Gamma}_{1:T}} E_{S_T}^{\gamma_{1:T}} \{u_T(s_T) \big| h_T\} \qquad (37)$$
$$s.t. \quad (14) - (16).$$

For all $t < T$,

$$V_t^*(h_t) = E_{A_t, Z_t, X_t, S_t}^{\sigma_t^*, \psi_t^*, \gamma_{1:t-1}^*} \Big\{ u_t(s_t, a_t) + \beta V_{t+1}^*\big(h_{t+1}\big) \big| h_t \Big\} \qquad (38)$$
$$s.t. \quad (32) - (34),$$

where $(\sigma_t^*, \psi_t^*, \gamma_{1:t-1}^*)$ satisfies

$$(\sigma_t^*, \psi_t^*, \gamma_{1:t-1}^*) = \arg \max_{\sigma_t \in \boldsymbol{\Sigma}_t} \min_{\psi_t \in \boldsymbol{\Psi}_t(h_t)} \min_{\gamma_{1:t-1} \in \boldsymbol{\Gamma}_{1:t-1}} E_{A_t, Z_{1:t}, X_{1:t}}^{\sigma_t, \psi_t, \gamma_{1:t-1}} \Big\{ \sum_{\tau = 1, 2, \ldots, t} \beta^\tau u_\tau(s_\tau, a_\tau) + \beta^{t+1} V_{t+1}^*\big(h_{t+1}\big) \big| h_t \Big\} \qquad (39)$$
$$s.t. \quad (32) - (34).$$

In (38) and (39), compared to (35) and (36), the maximin is only over σ_t and ψ_t instead of $\sigma_{t:T}$ and $\psi_{t:T}$. Moreover, the system dynamics constraints are only considered from time 1 to t (instead of 1 to T).

The dynamic programming equations developed in this section are complex, mainly because they require considering the system dynamics in all the previous periods. Moreover, V_t^* is a function of history h_t which has a dimensionality that grows in time. To tackle this complexity, we consider the special case of S-rectangular RPOMDPs in the next section, and introduce a sufficient statistic and an information state which have a time-invariant dimensionality. This significantly simplifies the dynamic programming equations, and allows considering system dynamics only at the current period.

5. Sufficient Statistic and Information State

We now develop a sufficient statistic that enables predicting the future behavior of RPOMDPs as well as an information state that can be used for their optimal control. To this end, we first extend the existing notions of sufficient statistic and information state for uncertain input-output systems (see, e.g., Bertsekas (1995)) to ambiguous input-output systems. Next, we use S-rectangularity—a stronger notion of rectangularity (see Wiesemann et al. (2013) and Online Appendix B)—to develop a sufficient statistic and an information state for RPOMDPs.

Definition 15 (**Sufficient Statistic for Ambiguous Random Variables**). Statistic θ is said to be sufficient for the ambiguous random variable I with respect to information h, if (a) it can be calculated as a function of h_t, and (b)the conditional ambiguity set of I given the statistic θ is independent of h:

$$\mathbf{\Gamma}(I \mid \theta, h) = \mathbf{\Gamma}(I \mid \theta). \tag{40}$$

This means that, besides the information provided by the sufficient statistic θ, there is no additional information about the ambiguous random variable I in h .

Definition 16 (**Information State of Ambiguous Input-Output Systems**). Statistic $\theta_t \in \mathbf{\Theta}_t$ is said to be an information state for the optimal control of an ambiguous input-output system (with the DM's information $h_t \in \boldsymbol{H}_t$), if there exists an optimal control policy $\hat{\sigma}^*_{1:T}(\theta_t)$ where $\hat{\sigma_t}^* : \mathbf{\Theta}_t \to \Delta(\boldsymbol{A}_t)$ for all $t \in \boldsymbol{T}$.

If $\theta_t \in \mathbf{\Theta}_t$ is an information state, then the search for the optimal policy can be restricted to the policy space $\hat{\mathbf{\Sigma}}_{1:T} = \{\hat{\sigma}_t : \mathbf{\Theta}_t \to \Delta(\boldsymbol{A}_t), \forall t \in \boldsymbol{T}\}$ instead of $\mathbf{\Sigma}_{1:T} = \{\sigma_t : \boldsymbol{H}_t \to \Delta(\boldsymbol{A}_t), \forall t \in \boldsymbol{T}\}$. In other words, all the relevant information of h_t for optimal control of the ambiguous input-output system exists in θ_t. Moreover, there are two other desired properties for the information state:

(i) θ_t is only a function of h_t, and hence, can be written as

$$\theta_t = r_t(h_t) \tag{41}$$

for some function r_t. This property ensures that the information state can be calculated with having access only to the DM's information at time t.

(ii) θ_t is a function of θ_{t-1} and the new information arriving between t and $t+1$, and hence, can be written as

$$\theta_t = q_t(\theta_{t-1}, h_t \backslash h_{t-1}) \tag{42}$$

for some function q_t. This property means that the DM does not need to record h_t and only needs to carry on θ_t.

Note that in the definitions of sufficient statistic and information state above, the static θ_t can include ambiguity sets. Moreover, from the above definitions, an information state provides adequate information for (a) evaluating the sum of the DM's past and future expected utilities for a given policy, and (b) determining the optimal policy. However, a sufficient statistic is only useful for predicting the DM's expected future utility. Furthermore, it should be noted that the information history h_t itself serves as both a sufficient statistic and an information state. However, as noted earlier, h_t grows in time, and thus, it is desirable to find alternatives that are time-invariant in size. We provide such a sufficient statistic and information state in the next sections.

5.1. Sufficient Statistic

In POMDPs, it is known that the DM's belief about the system state serves both as a sufficient statistic and as an information state (see, e.g., Bertsekas (1995)). However, as we saw earlier (see, e.g., Example 2), the story is a bit more nuanced for RPOMDPs: the existence of ambiguity in RPOMDPs changes this known result. In particular, in what follows, we show that while the belief about the system state (or the set of such beliefs when it is not unique) can still be used as a sufficient statistic in RPOMDPs (under certain sufficient conditions such as S-rectangularity), it requires some "augmentation" to serve as an information state.

THEOREM 5 **(Sufficient Statistic for S-rectangular RPOMDPs)**. *For an S-rectangular RPOMDP (see Definition 11), the set of possible conditional state distributions $\mathbf{\Pi}_t(h_t)$ defined as*

$$\mathbf{\Pi}_t(h_t) = \cup_{\gamma_{1:t-1}\in\mathbf{\Gamma}_{1:t-1}}\{\pi_t(h_t,\gamma_{1:t-1})\} \tag{43}$$

is a sufficient statistic.

EXAMPLE 4 (**Sufficient Statistic**). Consider the setup in Example 2, but assume that the physician does not know the state distribution of the arriving patient, π_1, and his belief is that $\pi_1 \in \mathbf{\Pi}_1 = \big\{(\eta, 1-\eta) : 0 \leq \eta \leq 1\big\}$. If the physician treats at $t = 1$, then his information history at time $t = 2$ is $h_2 = (\mathbf{\Pi}_1, tr, y_1)$. Considering the state dynamics in Figure 3, Bayesian updating of the DM's belief about the system state yields $\pi_2 \in \mathbf{\Pi}_2(h_2) = \Bigg\{\Bigg(\eta\times 1 + (1-\eta)\times 0.5, (1-\eta)\times 0.5\Bigg) : 0 \leq \eta \leq 1\Bigg\}$. The set of distributions $\mathbf{\Pi}_2(h_2)$ is a sufficient statistic at time $t = 2$ for predicting the future dynamics of the system.

5.2. Information State

We now provide an information state for RPOMDPs. This entails "augmenting" the sufficient statistic presented in Theorem 5 by including the reward-to-reach-values.

THEOREM 6 **(Information State for S-rectangular RPOMDP)**. *For an S-rectangular RPOMDP (see Definition 11), the set of possible conditional distributions of the system state along with the reward-to-reach values forms an information state. That is,*

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_t = \Bigg\{&(\pi_t, R_t) : \exists\gamma_{1:t-1}\in\mathbf{\Gamma}_{1:t-1} \quad s.t. \quad \pi_t(h_t,\gamma_{1:t-1}) = \pi_t \\ &\& \quad R_t = E_{X_{1:t-1},Z_{1:t-1}}\{\sum_{\tau=1,..,t-1}\beta^\tau u_\tau(s_\tau,a_\tau)|h_t,\gamma_{1:t-1}\}\Bigg\}\end{aligned} \tag{44}$$

is an information state. Also, if the same distribution π_t can be reached with more than one reward-to-reach value in the set $\tilde{\boldsymbol{\theta}}_t$, then refining $\tilde{\boldsymbol{\theta}}_t$ by only recording the minimum of such reward-to-reach values will result in a new set $\boldsymbol{\theta}_t$ which is still an information state.

We denote the space of the refined information state $\boldsymbol{\theta}_t$ by $\boldsymbol{\Theta}_t$.[19]

EXAMPLE 5 (**Information State**). Consider the setup in Example 4. At time $t=2$, the reward-to-reach value corresponding to $\pi_1=(\eta,1-\eta)$ is $\eta\times(-50)+(1-\eta)\times(-20)$. Augmenting the DM's set of beliefs by attaching the reward-to-reach value to each possible distribution of the system state at time $t=2$ forms the information state $\tilde{\boldsymbol{\theta}}_2(h_2)$:

$$\tilde{\boldsymbol{\theta}}_2(h_2)=\left\{\Big(\big(\eta\times 1+(1-\eta)\times 0.5,(1-\eta)\times 0.5\big),\eta\times(-50)+(1-\eta)\times(-20)\Big):0\le\eta\le 1\right\}. \quad (45)$$

Since in this example every distribution π_2 appears with a single utility in $\tilde{\boldsymbol{\theta}}_2(h_2)$, the refined set $\boldsymbol{\theta}_2(h_2)$ is the same as $\tilde{\boldsymbol{\theta}}_2(h_2)$. The space of the refined information state is $\boldsymbol{\Theta}_t=\big\{\big((\alpha,1-\alpha),r\big):0\le\alpha\le 1,r\in\mathbb{R}\big\}$.

The information state introduced in Theorem 6 reduces to (a) a single distribution in POMDPs and in RPOMDPs with a perfect DM, and (b) a single fully known state (i.e., a degenerate distribution) in RMDPs. That is, since in such special cases of RPOMDPs the reward-to-reach is not required to be recorded for different possible state distributions, it can be omitted from the information state.

Next, we develop dynamic programming equations for RPOMDPs using information state $\boldsymbol{\theta}_t$. To this end, we first introduce the space of DM's policies as a function of the information state. Consider a policy $\hat{\sigma}_{1:T}$ such that

$$\hat{\sigma}_\tau:\boldsymbol{\Theta}_\tau\to\Delta(\boldsymbol{A}_\tau)\quad\forall\tau=1,2,...,T. \quad (46)$$

We denote the set of all such policies by $\hat{\boldsymbol{\Sigma}}_{t:T}$. We also denote the ambiguity set of future random variables from the DM's perspective by $\boldsymbol{\Gamma}(Z_{t:T}|\boldsymbol{\theta}_t)$ for all $t=1,2,...,T$ and DM's information state $\boldsymbol{\theta}_t$. The set of possible nature's policies for choosing the distributions from this set is denoted by $\boldsymbol{\Psi}_{t:T}(\boldsymbol{\Theta}_t)$.

THEOREM 7 (**Dynamic Programming with Information State**). *For an S-rectangular RPOMDP, the optimal expected future utility determined by (36) can be calculated dynamically as a function of $\boldsymbol{\theta}_t$, the refined information state of Theorem 6, as follows. At time T,*

$$V_T^*(\boldsymbol{\theta}_T)=\min_{(\pi_T,R_T)\in\boldsymbol{\theta}_T}E_{\pi_t}\{u_T(s_T)\}. \quad (47)$$

For $t=1,2,...,T-1$, first set $\boldsymbol{\theta}_{t+1}=q(\boldsymbol{\theta}_t,a_t,y_t)$ and determine

$$(\hat{\sigma}_t^*,\psi_t^*,\pi_t^*)=\arg\max_{\hat{\sigma}_t\in\hat{\boldsymbol{\Sigma}}_t}\min_{\psi_t\in\boldsymbol{\Psi}_t(\boldsymbol{\theta}_t)}\min_{(\pi_t,R_t)\in\boldsymbol{\theta}_t}\left\{R_t+\beta E^{\hat{\sigma}_t,\psi_t,\pi_t}_{A_t,S_t,Z_t,X_t}\left[u_t(s_t,a_t)+\beta V^*_{t+1}\big(\boldsymbol{\theta}_{t+1}\big)\big|\boldsymbol{\theta}_t,\right]\right\} \quad (48)$$

[19] $\boldsymbol{\Theta}_t$ is a set of sets.

$$
\begin{aligned}
s.t. \quad & y_t^n = g_t^n(s_t, a_t, z_{t-1}, x_t), && (49)\\
& y_t = g_t(s_t, a_t, \gamma_t, w_t), && (50)\\
& s_{t+1} = f_t(s_t, a_t, v_t). && (51)
\end{aligned}
$$

Then, $V_t^*(\boldsymbol{\theta}_t)$ *satisfies*

$$
\begin{aligned}
& V_t^*(\boldsymbol{\theta}_t) = E_{A_t,S_t,Z_t,X_t}^{\hat{\sigma}_t^*,\psi_t^*,\pi_t^*}\left[u_t(s_t,a_t) + \beta V_{t+1}^*(\boldsymbol{\theta}_{t+1}) \,\middle|\, \boldsymbol{\theta}_t\right] \qquad (52)\\
& s.t. \quad (49) - (51).
\end{aligned}
$$

Comparing (52) with (38), it can be seen that the system dynamic constraints are only considered at time t (instead of 1 to t). Also, the size of $\boldsymbol{\Theta}_t$ is limited to $|\boldsymbol{\Gamma}_{1:T}|$ while the size of $\boldsymbol{H}_t$ grows as t increases.

Theorem 7 enables us to solve RPOMDPs in a relatively efficient way by using a dynamic programming formulation that only carries θ_t as opposed to h_t. In the next section, we provide a few other ways of reducing the computational complexity involved in solving RPOMDPs via dynamic programming.

6. Complexity Reduction

Solving RPOMDPs is inherently computationally difficult. This difficulty is mainly due to the complexity of both the policy space and the information state space. To reduce the complexity of the policy space, we find conditions under which deterministic policies are optimal for the DM.[20] To reduce the complexity of the information state space, we look for conditions under which the ambiguity set can be reduced to a smaller set. To do so, we define (a) the convex hull of an ambiguity set, and (b) the set of vertices of the convex hull of an ambiguity set. Then we prove a *convex hull equivalence* theorem, which helps reducing the size of the DM's ambiguity set to its vertices when the DM is absolutely imperfect. Finally, we provide a result to ensure that the set of vertices of a rectangular ambiguity set remains rectangular, and therefore, dynamic programming equations developed in the previous sections remain valid.

6.1. Policy Space Complexity

In general, the DM's policy space in RPOMDPs includes randomized policies (see, e.g., (9)). We prove that when nature is perfect adversarial (see Definition 14), there exists an optimal deterministic policy (see, e.g., Case 1 of Example 3 in Section 3.4). Therefore, in this special case,

[20] Since the space of deterministic policies is much less complex than that of randomized policies, this in turn allows restricting attention to a less complex space (when possible).

one can reduce the policy space in (9) to the space of the deterministic policies. A deterministic policy is of the form $\sigma_{1:T}$ such that

$$\sigma_t : \boldsymbol{H}_t \rightarrow \boldsymbol{A}_t, t = 1, 2, ..., T. \tag{53}$$

Theorem 8 **(Optimality of Deterministic Policies)**. *An RPOMDP with perfect adversarial nature has an optimal deterministic policy.*

It should be noted, however, that the above theorem may not hold when the nature is imperfect adversarial (see Example 3, Case 2 in Section 3.4).

6.2. Information State Space Complexity

Example 5 shows that even in a simple setup, the information state can be complex and intractable, mainly because of the size of the ambiguity set. We provide a result that enables us to effectively reduce the size of the ambiguity set. To do so, we first introduce the following definition of the convex hull of an ambiguity set.

Definition 17 (**Convex Hull of Ambiguity Set**). Consider an ambiguity set $\boldsymbol{\Gamma}$. The convex hull of $\boldsymbol{\Gamma}$ denoted by $Conv(\boldsymbol{\Gamma})$ is the smallest set of distributions that consists of all the convex combinations of the distributions in $\boldsymbol{\Gamma}$. That is, $Conv(\boldsymbol{\Gamma})$ is the smallest set that satisfies

$$\forall \gamma_1, \gamma_2, ..., \gamma_N \in \boldsymbol{\Gamma}, \alpha_1, \alpha_2, ..., \alpha_N \in [0,1], \sum_{n=1,2,...,N} \alpha_n = 1 : \sum_{n=1,2,...,N} \alpha_n \gamma_n \in Conv(\boldsymbol{\Gamma}), \tag{54}$$

where $N = |\boldsymbol{\Gamma}|$.

Note that from the definition above, since $Conv(\boldsymbol{\Gamma})$ is the smallest of such sets, every member of the convex hull set can be written as the convex combination of a number of distributions in $\boldsymbol{\Gamma}$.

Example 6 (**Convex Hull of Ambiguity Set**). Consider the Ellsberg example (Ellsberg 1961) with some modifications. An urn contains 90 balls. Of these 90 balls, 30 balls are red (R) and the other 60 are either blue (B) or green (G). We further assume that the difference between the number of green and blue balls is at least 20. The ambiguity set in this example is $\{(P(R), P(B), P(G)) = (1/3, \alpha, 2/3 - \alpha) : (0 \leq \alpha \leq 2/9) \cup (4/9 \leq \alpha \leq 2/3)\}$. The convex hull of the ambiguity set is $\{(P(R), P(B), P(G)) = (1/3, \alpha, 2/3 - \alpha) : 0 \leq \alpha \leq 2/3\}$.

Consider two RPOMDPs both with an absolutely imperfect DM, and assume that these two RPOMDPs are the same except that they have two different ambiguity sets. The following theorem shows that if these two ambiguity sets have the same convex hull, then the two RPOMDPs have the same optimal policy. This result allows reducing the complexity of the state space of RPOMDPs with an absolutely imperfect DM by replacing their ambiguity sets with a smaller ambiguity set that has the same convex hull.

Theorem 9 **(Convex Hull Equivalence)**. *Consider two ambiguity sets,* $\mathbf{\Gamma}^1_{1:T}$ *and* $\mathbf{\Gamma}^2_{1:T}$*. If* $Conv(\mathbf{\Gamma}^1_{1:T}) = Conv(\mathbf{\Gamma}^2_{1:T})$ *then the DM's optimal policy and optimal total utility for an RPOMDP with an absolutely imperfect DM under ambiguities* $\mathbf{\Gamma}^1_{1:T}$ *and* $\mathbf{\Gamma}^2_{1:T}$ *is the same. In addition, if* $Conv(\mathbf{\Gamma}^1_{1:T}) \subset Conv(\mathbf{\Gamma}^2_{1:T})$*, then the DM's optimal total utility under* $\mathbf{\Gamma}^1_{1:T}$ *is not lower than that under* $\mathbf{\Gamma}^2_{1:T}$*.*

Theorem 9 allows us to reduce the complexity of calculating the optimal policy of an RPOMDP (with an absolutely imperfect DM) by considering the alternative rectangular set $\mathbf{\Gamma}'_{1:T}$ that has two properties: (a) it has the same convex hull as $\mathbf{\Gamma}_{1:T}$, and (b) it has a much smaller cardinality than $\mathbf{\Gamma}_{1:T}$. The smallest of such choices for $\mathbf{\Gamma}'_{1:T}$ is the set of vertices of the convex hull of $\mathbf{\Gamma}_{1:T}$ defined below.

Definition 18 (**Ambiguity Set Vertices**). The set of vertices of the the ambiguity set $\mathbf{\Gamma}$ is the set $Vert(\mathbf{\Gamma})$ such that (i) $Vert(\mathbf{\Gamma}) \subset \mathbf{\Gamma}$, (ii) $Conv(Vert(\mathbf{\Gamma})) = Conv(\mathbf{\Gamma})$, and (iii) no distribution $\gamma \in Vert(\mathbf{\Gamma})$ can be written as the convex combination of two other members of $\mathbf{\Gamma}$.

Example 7 (**Ambiguity Set Vertices**). Consider the modified Ellsberg example described in Example 6. The set of vertices of the convex hull is $\{(1/3, 0, 2/3), (1/3, 2/3, 0)\}$.

Note that, as the above example shows, the reduction of an ambiguity set to its vertices is especially effective when the majority of the distributions in the ambiguity set are a convex combination of the other distributions in the set.

Finally, in order to use dynamic programming, we need to ensure that $Vert(\mathbf{\Gamma})$ remains rectangular when $\mathbf{\Gamma}$ is rectangular. We prove this in the following theorem.

Theorem 10 **(Rectangularity of the Set of Vertices)**. *If an RPOMDP is rectangular under the ambiguity set* $\mathbf{\Gamma}_{1:T-1}$*, then it is also rectangular under the ambiguity set* $Vert(\mathbf{\Gamma}_{1:T-1})$*. Furthermore, if ambiguous random variables* $Z_{1:T-1}$ *are strongly independent under an ambiguity set* $\mathbf{\Gamma}_{1:T-1}$*, then they are also strongly independent under the ambiguity set* $Vert(\mathbf{\Gamma}_{1:T-1})$*.*

The implication of Theorem 10 is as follows. If one can use dynamic programming for finding the optimal policy of an RPOMDP with ambiguity set $\mathbf{\Gamma}_{1:T-1}$, then s/he can also do so using the ambiguity set $Vert(\mathbf{\Gamma}_{1:T-1})$. Using the latter, however, is far more convenient since it can significantly reduce the underlying computational complexity.

7. Stochastic Zero-Sum Games with Imperfect Private Monitoring

Decision-making under ambiguity with a maximin criterion can typically be viewed as a zero-sum game between the DM and nature (see, e.g., Iyengar (2005)). For any given S-rectangular RPOMDP (see Definition 11), one can construct an equivalent stochastic zero-sum game with imperfect private monitoring and incomplete information. This game is equivalent to the RPOMDP in that, in the perfect Bayesian equilibrium of the game, one of the players has the same utility as the DM's optimal utility in the RPOMDP. Similarly, the perfect Bayesian equilibrium strategy of the two players in this game corresponds to the DM's and nature's optimal policies in the RPOMDP. This equivalence is general enough and holds for various DM's and nature's categories discussed in Sections 3.3 and 3.4. We state the main results here, and refer the reader to Online Appendix C for further details regarding the construction of the game.

Theorem 11 **(RPOMDP as a Stochastic Zero-Sum Game)**. *Consider an S-rectangular RPOMDP with the DM's optimal policy $\sigma^*_{1:T-1}$ resulting in optimal value V^* under nature's policy $\psi^*_{1:T-1}$. There exists a corresponding zero-sum game with imperfect private monitoring and incomplete information that has a perfect Bayesian equilibrium with player one's strategy being $\tilde{\sigma}^*_{1:T} = \sigma^*_{1:T}$ and player two's strategy being $\tilde{\psi}^*_{1:T} = \psi^*_{1:T}$. In this game, player one's total utility from time 1 to T is $\tilde{V}^* = V^*$, where $\sigma^*_{1:T}, \psi^*_{1:T}$ and V^* can be calculated using (35) and dynamic programming equations of Theorem 4.*

8. Conclusions

We present a new dynamic decision-making framework termed Robust Partially Observable Markov Decision Process (RPOMDP), which allows the decision-maker to face (a) ambiguity with respect to system transitions, and (b) imperfect state observation. This new framework allows making robust decisions in a variety of applications by reducing the decision-maker's reliance on a single probabilistic model of transitions.

We establish conditions under which RPOMDPs can be solved via dynamic programming. We also provide a sufficient statistic and an information state for RPOMDPs that simplify the dynamic programming equations, and highlight various important differences between RPOMDPs, POMDPs, and RMDPs.

We also discuss ways in which the underlying computational complexities in RPOMDPs can be reduced. We do so by providing results that allow reducing the complexities of both the policy space (via considering deterministic policies) and the information space (via considering a smaller ambiguity set). We also build an important bridge between RPOMDPs and stochastic zero-sum games with imperfect private monitoring and incomplete state information, and note that RPOMDPs can be used for analyzing the perfect Bayesian equilibrium of such games.

We leave it to future research to provide other mechanisms for analyzing RPOMDPs. This can be done, for example, by considering approximation techniques, providing structural results that can facilitate the search for an optimal policy, and/or by utilizing results from stochastic games. Future research can also examine properties of RPOMDPs in infinite-horizon, and/or extend the results of this paper to cases with continuous state, action, or observation space. Finally, given various applications in which a decision maker faces both imperfect observations and ambiguity in transition probabilities, future research can shed light on the performance of RPOMDPs in a variety of real-world settings.

Acknowledgments

The work of the second author was partially supported by the National Science Foundation (NSF) under the Award Number CMMI-1562645.

References

Altman, E. and Solan, E. (2009), 'Constrained games: The impact of the attitude to adversary's constraints', *IEEE Transactions on Automatic Control* **54**(10), 2435–2440.

Bäauerle, N. and Rieder, U. (2017), 'Partially observable risk-sensitive Markov decision processes', *Mathematics of Operations Research* **42**(4), 1180–1196.

Başar, T. and Bernhard, P. (2008), *H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Springer Science & Business Media, New York, NY.

Bäuerle, N. and Rieder, U. (2013), 'More risk-sensitive Markov decision processes', *Mathematics of Operations Research* **39**(1), 105–120.

Bernhard, P. (2000), 'Max-plus algebra and mathematical fear in dynamic optimization', *Set-Valued Analysis* **8**(1-2), 71–84.

Bertsekas, D. (1995), *Dynamic Programming and Optimal Control*, Vol. 1, Athena Scientific, Belmont, MA.

Bertsekas, D. and Rhodes, I. (1973), 'Sufficiently informative functions and the minimax feedback control of uncertain dynamic systems', *IEEE Transactions on Automatic Control* **18**(2), 117–124.

Ellsberg, D. (1961), 'Risk, ambiguity, and the savage axioms', *The Quarterly Journal of Economics* **75**(4), 643–669.

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

Ghirardato, P., Maccheroni, F. and Marinacci, M. (2004), 'Differentiating ambiguity and ambiguity attitude', *Journal of Economic Theory* **118**(2), 133–173.

Gilboa, I. and Schmeidler, D. (1989), 'Maxmin expected utility with non-unique prior', *Journal of Mathematical Economics* **18**(2), 141–153.

González-Trejo, J., Hernández-Lerma, O. and Hoyos-Reyes, L. F. (2002), 'Minimax control of discrete-time stochastic systems', *SIAM Journal on Control and Optimization* **41**(5), 1626–1659.

Hansen, L. P. and Sargent, T. J. (2005), 'Robust estimation and control under commitment', *Journal of Economic Theory* **124**(2), 258–301.

Hansen, L. P. and Sargent, T. J. (2007), 'Recursive robust estimation and control without commitment', *Journal of Economic Theory* **136**(1), 1–27.

Itoh, H. and Nakamura, K. (2007), 'Partially observable Markov decision processes with imprecise parameters', *Artificial Intelligence* **171**(8), 453–490.

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

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

Klibanoff, P., Marinacci, M. and Mukerji, S. (2009), 'Recursive smooth ambiguity preferences', *Journal of Economic Theory* **144**(3), 930–976.

Maccheroni, F., Marinacci, M. and Rustichini, A. (2006), 'Dynamic variational preferences', *Journal of Economic Theory* **128**(1), 4–44.

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

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

Rasouli, M., Miehling, E. and Teneketzis, D. (2018), A scalable decomposition method for the dynamic defense of cyber networks, *in* ‘Game Theory for Security and Risk Management’, Springer, Aktiengesellschaft, Switzerland.

Saghafian, S. (2015), ‘Ambiguous partially observable Markov decision processes: Structural results and applications’, *Working Paper, Harvard University (SSRN 2508776)* .

Saghafian, S. and Tomlin, B. (2016), ‘The newsvendor under demand ambiguity: Combining data with moment and tail information’, *Operations Research* **64**(1), 167–185.

Satia, J. K. and Lave, R. E. (1973), ‘Markovian decision processes with uncertain transition probabilities’, *Operations Research* **21**(3), 728–740.

Solan, E. and Ziliotto, B. (2016), Stochastic games with signals, *in* ‘Advances in Dynamic and Evolutionary Games’, Springer, Aktiengesellschaft, Switzerland, pp. 77–94.

Strzalecki, T. (2011), ‘Axiomatic foundations of multiplier preferences’, *Econometrica* **79**(1), 47–73.

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

Witsenhausen, H. S. (1966), Minimax control of uncertain systems, Technical report, ESL-R-269, Electronic Systems Laboratory, MIT, Cambridge, MA.