Imprecise Probabilities Meet Partial Observability: Game Semantics for Robust POMDPs

Eline M. Bovy[1] , **Marnix Suilen**[1] , **Sebastian Junges**[1] and **Nils Jansen**[1,2]
[1]Radboud University, The Netherlands
[2]Ruhr-University Bochum, Germany
{eline.bovy, marnix.suilen, sebastian.junges}@ru.nl, n.jansen@rub.de

Abstract

Partially observable Markov decision processes (POMDPs) rely on the key assumption that probability distributions are precisely known. Robust POMDPs (RPOMDPs) alleviate this concern by defining imprecise probabilities, referred to as uncertainty sets. While robust MDPs have been studied extensively, work on RPOMDPs is limited and primarily focuses on algorithmic solution methods. We expand the theoretical understanding of RPOMDPs by showing that 1) different assumptions on the uncertainty sets affect optimal policies and values; 2) RPOMDPs have a partially observable stochastic game (POSG) semantic; and 3) the same RPOMDP with different assumptions leads to semantically different POSGs and, thus, different policies and values. These novel semantics for RPOMDPs give access to results for POSGs, studied in game theory; concretely, we show the existence of a Nash equilibrium. Finally, we classify the existing RPOMDP literature using our semantics, clarifying under which uncertainty assumptions these existing works operate.

1 Introduction

Partially observable Markov decision processes (POMDPs) are the standard model for decision-making under stochastic uncertainty and incomplete state information [Kaelbling *et al.*, 1998]. A common objective in a POMDP is for an agent to compute a policy that maximizes the expected discounted reward. While POMDPs have been studied extensively, a key assumption planning methods for POMDPs rely on is that the model dynamics, *i.e.*, the transition and observation probabilities, are precisely known. Under that assumption, it is known that an optimal policy of a POMDP is the solution to a fully observable infinite-state *belief* MDP [Kaelbling *et al.*, 1998].

In the fully observable setting, Markov decision processes (MDPs) [Puterman, 1994] have been extended to *robust* MDPs (RMDPs) to account for an additional layer of uncertainty around the probabilities that govern the model dynamics known as the uncertainty set. These RMDPs have been studied extensively, in terms of their semantics [Iyengar, 2005, Nilim and Ghaoui, 2005, Wiesemann *et al.*, 2013], efficient algorithms to solve specific classes of RMDPs [Behzadian *et al.*, 2021, Ho *et al.*, 2021, Wang *et al.*, 2023], and their application in reinforcement learning [Jaksch *et al.*, 2010, Petrik and Subramanian, 2014, Suilen *et al.*, 2022, Moos *et al.*, 2022].

Robust MDPs can be seen as games between the *agent*, who aims to maximize their reward by choosing an action at each state, and *nature*, who aims to minimize the agent's reward by selecting adversarial probability distributions from the uncertainty set. As a consequence, RMDPs and zero-sum stochastic games (SG) [Shapley, 1953, Gillette, 1957] are closely related, see in particular [Iyengar, 2005, Section 5] for a reduction from (finite horizon) RMDP to SG.

For RMDPs, two semantics exist for nature's behavior when encountering the same state and action twice. *Static* uncertainty semantics require nature to always select the same probability distribution, while *dynamic* uncertainty semantics allow nature to make a new choice every time a state-action pair is encountered. [Iyengar, 2005, Lemma 3.3] established that for finite horizon and discounted infinite horizon reward maximization in certain RMDPs, static and dynamic uncertainty semantics coincide, meaning that for a given agent's policy, both semantics result in precisely the same value.

Extensions to robust POMDPs (RPOMDPs) exist [Osogami, 2015, Chamie and Mostafa, 2018, Saghafian, 2018, Suilen *et al.*, 2020, Nakao *et al.*, 2021, Cubuktepe *et al.*, 2021, Bovy, 2023], but primarily focus on algorithmic approaches to compute optimal policies. Notably, these algorithms compute optimal policies under different implicit assumptions on the semantics of RPOMDPs, particularly concerning static and dynamic uncertainty.

Contributions. This paper sets out to clarify and expand the theoretical understanding of RPOMDPs. Specifically, we define semantics with associated value functions and policies for RPOMDPs under various assumptions on the uncertainty. We explicitly define the semantics of RPOMDPs via zero-sum two-sided partially observable stochastic games (POSGs) [Delage *et al.*, 2023]. Our key contributions are:

1. **Uncertainty assumptions matter.** We introduce a continuum of uncertainty assumptions for RPOMDPs called *stickiness*. Stickiness determines when nature's choices for resolving the uncertainty become fixed. The two extremes, immediately and never, coincide with the static

and dynamic uncertainty semantics of RMDPs. We show in Theorem 1 that, in contrast to RMDPs, these two extremes no longer coincide for RPOMDPs. Specifically, they may lead to different optimal values. Moreover, the *order of play* (whether the agent or nature makes the first move) matters. We show that the differences in these assumptions can lead to significant differences in optimal values. We account for these results by providing a new RPOMDP definition that explicitly accounts for these uncertainty assumptions in Definition 3.

2. **Robust POMDPs are POSGs.** We provide a formal POSG semantic for RPOMDPs with explicit stickiness and order of play. We establish a direct correspondence between policies of POSGs and RPOMDPs that ensure equal values for both models (Theorem 2). Moreover, different uncertainty assumptions in the RPOMDP lead to semantically different POSGs and hence explain the result listed in Contribution 1. Finally, we use the POSG semantics to prove the existence of Nash equilibria, which we use in turn to prove the existence of optimal values in finite horizon RPOMDPs (Theorem 3).
3. **Classification of existing RPOMDP works.** We provide a classification of existing RPOMDP literature into our semantic framework (Section 5).

The extended version of this paper, with all the appendices, can be found at [Bovy *et al.*, 2024].

2 Preliminaries

A discrete probability distribution over a finite set X is a function $\mu\colon X \to [0,1]$ such that $\sum_{x\in X}\mu(x) = 1$. For infinite sets, we only consider finite probability distributions. That is, for an infinite set X, a finite probability distribution over X is a function $\lambda\colon X \to [0,1]$ with finitely many $x \in X.\ \lambda(x) \neq 0$ and $\sum_{x\in X}\lambda(x) = 1$. The set of all probability distributions over X is denoted as $\Delta(X)$, and $\mathcal{P}(X)$ is the powerset of X. By $(X \to Y)$, we denote the set of all functions $f\colon X \to Y$, and $f\colon X \hookrightarrow Y$ for a partial function. The symbol $\perp$ is used for undefined. Finally, we use Currying to describe functions that map to functions, *e.g.*, $f\colon X \to (Y \to Z)$ represents a function that maps each $x \in X$ to a function $g_x\colon Y \to Z$.

2.1 Markov Models

Definition 1 (POMDP)**.** *A partially observable Markov decision process (POMDP) is a tuple* $\langle S, A, T, R, Z, O\rangle$ *where* S, A, Z *are finite sets of states, actions, and observations, respectively.* $T\colon S\times A \to \Delta(S)$, $R\colon S\times A\to\mathbb{R}$, *and* $O\colon S \to Z$ *are the transition, reward, and observation functions, respectively.*

This definition uses POMDPs with *deterministic observations*, in contrast to the more standard stochastic observation functions [Kaelbling *et al.*, 1998]. However, every POMDP with stochastic observations can be transformed into such a POMDP [Chatterjee *et al.*, 2016]. For convenience, we sometimes write $T(s,a,s')$ for $T(s,a)(s')$.

A *Markov decision process* (MDP) is a POMDP where all states are fully observable. We simplify the tuple definition to $\langle S, A, T, R\rangle$ in the MDP case.

Paths and histories. A *path* in a (PO)MDP M is a sequence of successive states and actions: $\tau = \langle s_0, a_0, \ldots, s_n\rangle \in (S\times A)^*\times S$ such that $T(s_i, a_i, s_{i+1}) > 0$ for all $i \geq 0$. We denote the set of paths in M by Paths^M. The concatenation of two paths is written as $\tau \oplus \tau'$. A history in a POMDP is a sequence of observations and actions observed from a path $\langle s_0, a_0, \ldots\rangle$: $h \in (Z\times A)^*\times Z$ such that $h = \langle O(s_0), a_0, O(s_1), a_1 \ldots\rangle$.

Policies. A history-based *stochastic* policy[1] is a function that maps histories to distributions over actions, that is, $\pi\colon (Z\times A)^*\times Z \to \Delta(A)$. The policy π is *deterministic*, or pure, if it only maps to single actions, and stationary if its domain is Z, *i.e.*, it only maps the current observation. The set of all history-based stochastic policies is denoted by Π and the set of all history-based deterministic policies by Π^{det}. A history-based *mixed* policy is a probability distribution over the set of history-based deterministic policies, that is, $\pi^{mix} \in \Delta(\Pi^{det})$ The set of all history-based mixed policies is denoted by Π^{mix}. Throughout the rest of the text, unless otherwise mentioned, all policies are history-based , and unless indicated by either det or mix, the (sets of) policies are stochastic.

Values. We maximize the expected reward, either with a finite horizon $K \in \mathbb{N}$ (denoted fh) or in the infinite horizon with a discount factor $\gamma \in (0,1)$ (denoted dih). We denote these *objectives* by $\phi \in \{\text{fh}, \text{dih}\}$. The value of a policy $\pi \in \Pi$ in a (PO)MDP for the objective ϕ is given by the value function V^π_ϕ, and the optimal value is V^*_ϕ. The value of a policy for either objective is [Spaan, 2012]:

$$V^\pi_{\text{fh}} = \mathbb{E}\left[\sum_{t=0}^{K-1} r_t \mid \pi\right], \quad V^\pi_{\text{dih}} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right],$$

where r_t is the reward collected at time t under policy π. The optimal value V^*_ϕ is defined as $\sup_{\pi\in\Pi} V^\pi_\phi$.

2.2 Robust MDPs

Robust MDPs extend standard MDPs by defining an uncertainty set of probability distributions that a state-action pair can map to instead of a single fixed and known distribution. Let U be a finite set of variables, and define $\boldsymbol{U} \subseteq (U \to \mathbb{R})$ as the uncertainty set. Let $\boldsymbol{U}$ be non-empty, a robust MDP is then defined as follows.

Definition 2 (RMDP)**.** *A robust MDP (RMDP) is a tuple* $\langle S, A, \boldsymbol{T}, R\rangle$ *where* S, A, *and* R *are again states, actions, and the reward function.* $\boldsymbol{T}\colon \boldsymbol{U} \to (S\times A \to \Delta(S))$ *is the uncertain transition function, consisting of a possibly infinite set of transition functions* $T\colon S\times A\to\Delta(S)$, *where every* $T \in \boldsymbol{T}$ *is determined by a variable assignment* $(U \to \mathbb{R}) \in \boldsymbol{U}$.

Remark 1. The variable assignment $\boldsymbol{U}$ maps the variables to $\mathbb{R}$ and not to $[0,1]$ as mapping to the reals gives more freedom in defining the uncertainty set, allowing for more complicated dependencies between transitions. The uncertain transition function $\boldsymbol{T}$ ensures that all state-action pairs are mapped to probability distributions.

[1] Also known as a behavioral strategy.

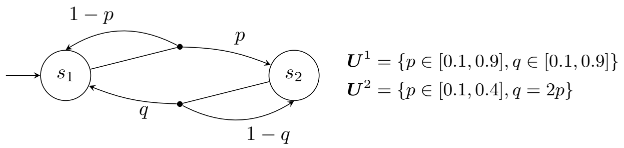

Figure 1: An example RMDP with two uncertainty sets.

Game interpretation. As already mentioned in the introduction, we interpret RMDPs as games between the agent, who selects actions through a policy $\pi : (S \times A)^* \times S \to \Delta(A)$, and nature, who uses its policy $\theta : (S \times A \times \boldsymbol{U})^* \times S \to \Delta(\boldsymbol{U})$ to select variable assignments $u \in \boldsymbol{U}$ from the uncertainty set to determine the probability distributions, such that $\boldsymbol{T}$ is non-empty. That is, any variable selection u must yield a valid probability distribution for all state-action pairs:

$$\forall s \in S, a \in A.\, \boldsymbol{T}(u)(s,a) \in \Delta(S).$$

The sets of the agent's and nature's policies are again Π and Θ, respectively. The sets of deterministic and mixed policies are constructed analogously as for POMDPs.

The maximal value that a policy can achieve over all possible ways to resolve the uncertainty is defined for both objectives, respectively, as

$$V^*_{\text{fh}} = \sup_{\pi \in \Pi} \inf_{\theta \in \Theta} \mathbb{E}\left[\sum_{t=0}^{K-1} r_t\right],\; V^*_{\text{dih}} = \sup_{\pi \in \Pi} \inf_{\theta \in \Theta} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right].$$

It is often assumed that nature plays stationary and deterministic in RMDPs. Under certain conditions on the uncertainty set, this assumption is non-restrictive as nature's best policy falls within this class [Iyengar, 2005, Wiesemann *et al.*, 2013, Grand-Clément *et al.*, 2023].

Remark 2. Our definition of RMDPs is more general than common definitions: Most RMDP definitions assume a form of independence in the uncertainty set between different states (or actions), known as s- (or (s,a)-) *rectangularity*. Our definition subsumes these rectangular RMDPs. While rectangular RMDPs satisfy a saddle point condition, meaning the $\sup \inf$ may be reversed in the definition of V^*_ϕ, this has not been shown for RMDPs in general. Our result in Theorem 3 shows that the saddle point condition holds for RPOMDPs in general for finite horizon. This extends to RMDPs using a fully observable observation function. We refer to [Wiesemann *et al.*, 2013] for a more standard definition of rectangularity and an overview of the computational properties of rectangular RMDPs, and [Jansen *et al.*, 2022] for an overview on non-rectangular RMDPs.

Example 1. Figure 1 depicts a small RMDP together with two possible uncertainty sets $\boldsymbol{U}^1$ and $\boldsymbol{U}^2$. In this RMDP, the agent only has singleton choices, while nature chooses variable assignments for p and q. Given an uncertainty set and a variable assignment in that uncertainty set, for example, $u = \{p \mapsto 0.3, q \mapsto 0.5\} \in \boldsymbol{U}^1$, we get a fully defined transition function. $\boldsymbol{U}^1$ is an (s,a)-rectangular uncertainty set since each variable influences the transition probabilities in only one state-action pair. $\boldsymbol{U}^1$ could hence be split into two independent uncertainty sets: $\boldsymbol{U}^1 = \{p \in [0.1, 0.9]\} \times \{q \in [0.1, 0.9]\}$. In contrast, $\boldsymbol{U}^2$ is not (s,a)-rectangular, since the value of q depends on p, so p influences transitions from state s_1 as well as from state s_2.

Static and dynamic uncertainty. A prominent semantic concern on RMDPs is whether nature must play consistently when a state is repeatedly visited. *Static uncertainty* semantics require nature to choose a single variable assignment $u \in \boldsymbol{U}$ once-and-for-all, fixing all probability distributions form the start. On the other hand, *dynamic uncertainty* semantics allow nature to choose a new variable assignment independently each time a state is visited. In [Iyengar, 2005, Lemma 3.3], it is shown that on (s,a)-rectangular RMDPs with a finite horizon or discounted infinite horizon objective, these semantics, and thus the values, coincide.

Remark 3. Although our use of variables in the transition function is similar to, *e.g.*, [Wiesemann *et al.*, 2013], it is not standard. Often, the transition function directly maps to uncertainty sets, *e.g.*, [Iyengar, 2005, Nilim and Ghaoui, 2005, Ho *et al.*, 2018]. The use of variables has the following benefits over directly mapping to uncertainty sets: (1) support for various semantics, such as different forms of rectangularity, without changing the signature of the uncertain transition function $\boldsymbol{T}$; (2) it allows us to keep track of partial restriction on nature's choice, which is needed when moving to the partially observable setting (Section 3.1).

3 RPOMDPs and Uncertainty Assumptions

In this section, we define a game-based framework for robust POMDP semantics that can be instantiated by making different *uncertainty assumptions*. Specifically, we incorporate two key assumptions into our RPOMDP definition: *stickiness* and *order of play*. Stickiness concerns the moment at which nature must choose the values of the variables U and extends static and dynamic uncertainty from RMDPs to the partially observable setting. The order of play specifies whether the agent or nature moves first. It determines the moment nature observes the most recent agent action.

This section is structured as follows. We briefly discuss our assumptions about partial observability to introduce notation needed and then formally define RPOMDPs. Next, we clarify how notions such as paths and histories carry over from POMDPs and RMDPs to RPOMDPs. We briefly describe the order-of-play assumption and provide a more elaborate discussion of stickiness in Section 3.1. Finally, in Section 3.2, we discuss the optimal value of RPOMDPs under different uncertainty assumptions and demonstrate that these assumptions matter, *i.e.*, yield different optimal values (Theorem 1).

RPOMDPs. Analogous to RMDPs, we interpret RPOMDPs as a game between the agent and nature. To make our RPOMDP definition as general as possible, we assume partial observability for both the agent and nature. We factorize the observations into three parts: *private* observations of agent and nature, respectively, and *public* observations that both players observe. Hence, each player obtains two observations in each state. For the remainder of the paper, we use $\mathfrak{a}$ and $\mathfrak{n}$ to denote whether a set or function belongs to the agent or to nature, respectively. Likewise, we use $\bullet$ and $\circ$ to denote whether a set or function relates to private or public observations.

Definition 3 (RPOMDP). *A robust POMDP (RPOMDP) is a tuple* $\langle S, A, \boldsymbol{T}, R, Z_\bullet^{\mathfrak{a}}, Z_\bullet^{\mathfrak{n}}, Z_\circ, O_\bullet^{\mathfrak{a}}, O_\bullet^{\mathfrak{n}}, O_\circ, \mathsf{stick}, \mathsf{play}\rangle$*, where* $S, A, \boldsymbol{T}$*, and* R *are sets of states and actions, the uncertain transition function, and the reward function, as in RMDPs. The sets* $Z_\bullet^{\mathfrak{a}}, Z_\bullet^{\mathfrak{n}}$*, and* $Z_\circ$ *are the private observations for the agent, for nature, and the public observations, respectively.* $O_\bullet^{\mathfrak{a}} \colon S \to Z_\bullet^{\mathfrak{a}}, O_\bullet^{\mathfrak{n}} \colon S \to Z_\bullet^{\mathfrak{n}}$*, and* $O_\circ \colon S \to Z_\circ$ *are the observation functions belonging to the agent, nature, and public observations.* $\mathsf{stick} \colon U \times Z_\bullet^{\mathfrak{n}} \times Z_\circ \times A \to \{0,1\}$ *is the* stickiness function*, and* $\mathsf{play} \in \{\mathfrak{a}, \mathfrak{n}\}$ *the* order of play*,* i.e.*, which player moves first.*

As for POMDPs, we consider deterministic observations. We show in Appendix B that RPOMDPs with stochastic or uncertain observations can be rewritten in RPOMDPs with deterministic observations.

Paths and histories. A path through an RPOMDP M is a sequence $\tau = \langle s_0, a_0, u_0, s_1, \ldots, s_n\rangle \in (S \times A \times \boldsymbol{U})^* \times S$ that consists of environment states, agent actions, and nature's variable assignments $u \in \boldsymbol{U}$, such that for all $i > 0$:

$$\boldsymbol{T}(u_{i-1})(s_{i-1}, a_{i-1}, s_i) > 0.$$

As before, we denote the set of paths in M by Paths^M. A history is the observable fragment of a path for either the agent or nature. The agent's histories are sequences in $H^{\mathfrak{a},M} \subseteq (Z_\bullet^{\mathfrak{a}} \times Z_\circ \times A)^* \times Z_\bullet^{\mathfrak{a}} \times Z_\circ$, observing the agent's private and public observations of the states and its own actions. Nature's histories are sequences in $H^{\mathfrak{n},M} \subseteq (Z_\bullet^{\mathfrak{n}} \times Z_\circ \times A \times \boldsymbol{U})^* \times Z_\bullet^{\mathfrak{n}} \times Z_\circ$, observing its private and public observations of the states, the agent's actions, and variable assignments $u \in \boldsymbol{U}$ that resolve the uncertainty. The histories for the agent and nature are obtained from a path by applying the relevant observation functions, respectively, similar to POMDPs. We give an explicit mapping in Appendix A.2.

Order of play. For any given path, both the agent and nature must make a move. We consider turn-based games and must, therefore, select who picks their move first[2]. We encode this information directly in the signature of the nature policy below. We remark that after both players have made their move, the resulting state is equivalent as we assume that nature always observes the actions picked previously.

Policies. As with RMDPs, we denote the agent's policies by $\pi \in \Pi$ and nature's by $\theta \in \Theta$. Specifically, the agent's policies are defined as maps from the agent's histories to distributions over actions $\pi \colon H^{\mathfrak{a},M} \to \Delta(A)$. Nature's policies are maps from nature's histories and the last agent action to *finite* distributions over variable assignments $\theta \colon H^{\mathfrak{n},M} \times A \to \Delta(\boldsymbol{U})$. When nature moves first, the last agent action is not available and therefore not part of nature's policy: $\theta \colon H^{\mathfrak{n},M} \to \Delta(\boldsymbol{U})$. The sets of deterministic and mixed policies are constructed analogously as for POMDPs in Section 2.

[2]In our setting, the case that both players pick their actions simultaneously is equivalent to letting nature move first, as we assume the agent never directly observes the selection of nature. See [Kwiatkowska *et al.*, 2022] for more information about simultaneous stochastic games.

3.1 Stickiness: Restricting Nature's Choices

Stickiness describes whether nature's choice at one point should remain fixed ('stick') in the future[3]. The simplest instances of stickiness are when nature's choices never stick or when they all stick from the start. If nature's choices never stick, so values never stick to variables, we say the RPOMDP has *zero* stickiness. If nature's choices stick from the start, so values directly stick to all variables, we say the RPOMDP has *full* stickiness. Zero and full stickiness correspond to dynamic and static uncertainty in RMDPs, respectively.

Zero and full stickiness are only the two extremes of a spectrum of different stickiness types. In addition, RPOMDPs admit partial types of stickiness, where nature may have to fix variable values but can delay some choices depending on the specific stickiness function. We now give an intuitive example on stickiness before moving to the formal definition. For explicit examples of stickiness, including so-called *observation-based* stickiness, see Appendix C.

Example 2 (Stickiness). Consider the following drone delivery problem, naturally modeled as (R)POMDP. The agent controls a drone that has to deliver packages. States encode the drone's location, actions are direction and speed adjustments, and observations are location estimations. The transition probabilities represent the chance of reaching adjacent locations. Different types of stickiness can model different sources of uncertainty on those probabilities:

Full stickiness. The drone experiences an unknown drift probability caused by, *e.g.*, a dented blade. The agent must account for this unknown but *fixed probability*.

Zero stickiness. Wind influences the probability of reaching adjacent states. While predictable to a certain degree, a margin of uncertainty will remain. As the wind changes over time, the agent has to account for *changing probabilities*.

Partial stickiness. We need partial stickiness when nature eventually has to commit to a probability, but not from the start. Suppose we extend our problem with a municipality that has created no-fly zones and will install monitors in these zones to detect violations. We encode the no-fly zones in the state space to reason about the probability of the agent being detected. Initially, the municipality will try out possible placements for their monitors. The probability of being detected, hence, lies in an uncertainty set formed by the different placements of monitors. Once the placement of the monitors is final, the probability of getting caught in a no-fly zone becomes fixed. A partial stickiness function that returns 1 when observing a drone in a no-fly zone, fixing the number of monitors at that point, captures such scenarios.

We allow partial stickiness to depend on what nature observes, *i.e.*, its private observations $Z_\bullet^{\mathfrak{n}}$, public observations $Z_\circ$, and the agent's actions A.

Definition 4. *The* stickiness *of an RPOMDP is a Boolean function indicating whether nature's choice of a value for variable* $v \in U$ *should remain fixed:*

$$\mathsf{stick} \colon U \times Z_\bullet^{\mathfrak{n}} \times Z_\circ \times A \to \{0,1\}.$$

[3]The name follows from the idea that nature always chooses values for all variables, but some values stick for the rest of time. Whether a variable sticks is determined by the stickiness.

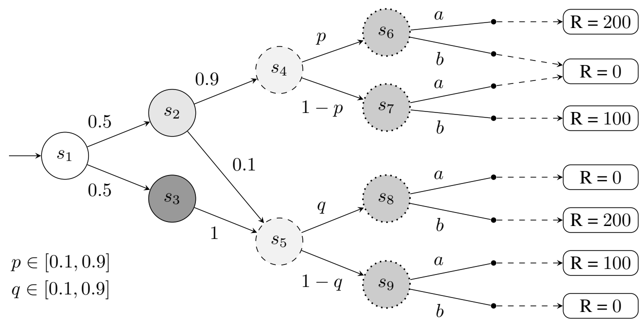

Figure 2: An RPOMDP where full and zero stickiness do not coincide in their optimal value.

Below, we describe how we use the stick function to compute restrictions on nature's choices and, with that, define valid nature policies.

Fixed variables and agreeing assignments. Depending on the stickiness of the RPOMDP, past choices of nature may restrict its future choices. Let $\boldsymbol{U}^{\hookrightarrow}$ denote the set of partial variable assignments $U \hookrightarrow \mathbb{R}$. Let $u^{\perp} \in \boldsymbol{U}^{\hookrightarrow}$ be the totally undefined variable assignment: $\forall v \in U. u^{\perp}(v) = \perp$. We define a function $\mathsf{fix}\colon \text{Paths}^M \to \boldsymbol{U}^{\hookrightarrow}$ such that $\mathsf{fix}(\tau)$ defines the partial variable assignment that remains fixed based on the stickiness function. This function is inductively defined as $\mathsf{fix}(s_I) = \emptyset = u^{\perp}$ for the initial path s_I, and

$$\mathsf{fix}(\tau \oplus \langle a, u, s' \rangle)(v) =$$
$$\begin{cases} u(v) & \text{if } \mathsf{fix}(\tau)(v) \text{ undefined, } v \in U^{\text{stick}}(last(\tau), a), \\ \mathsf{fix}(\tau)(v) & \text{otherwise,} \end{cases}$$

using $\quad U^{\text{stick}}(s, a) = \{v \mid \mathsf{stick}(v, O^{\mathtt{n}}_{\bullet}(s), O_{\circ}(s), a) = 1\}$

to denote the variables that stick. We can straightforwardly lift the definition of fix to nature's histories using

$$U^{\text{stick}}_h(z^{\mathtt{n}}_{\bullet}, z_{\circ}, a) = \{v \mid \mathsf{stick}(v, z^{\mathtt{n}}_{\bullet}, z_{\circ}, a) = 1\}.$$

Two partial functions agree if they assign equal values to all defined inputs. We use $\boldsymbol{U}^{\mathcal{P}}(u)$ for the variable assignments that agree with partial variable assignment u.

Valid paths, histories, and policies. Let $\tau = \langle s_0, a_0, u_0, s_1, \dots, s_n \rangle \in \text{Paths}^M$. For $k < n$, we denote the prefix $\tau_{0:k} = \langle s_0, a_0, u_0, s_1, \dots, s_k \rangle$. A path is valid, if for every $k < n$, $u_k \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(\tau_{0:k}))$. A history is valid if it corresponds to some valid path. A nature policy is valid if all variable assignments that nature randomizes over given a history and action are in the set of variable assignments that agree with the variable restrictions generated by the history. That is, $\forall h^{\mathtt{n}} \in H^{\mathtt{n}}, \forall a \in A, \forall u \in \boldsymbol{U}$.

$$\theta(h^{\mathtt{n}}, a)(u) > 0 \implies u \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(h^{\mathtt{n}})).$$

From here on, all paths, histories, and policies are assumed to be valid.

3.2 The Value of an RPOMDP

Values. The values of an RPOMDP given agent policy $\pi \in \Pi$ and nature policy $\theta \in \Theta$ for both the finite horizon

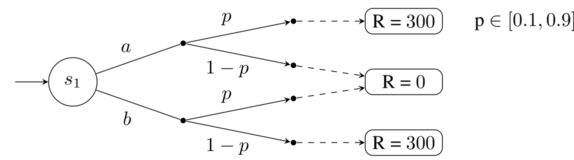

Figure 3: An RPOMDP where nature first and agent first semantics do not coincide in their optimal value.

and discounted infinite horizon objective are

$$V^{\pi,\theta}_{\text{fh}} = \mathbb{E}\left[\sum_{t=0}^{K-1} r_t \mid \pi, \theta\right], \quad V^{\pi,\theta}_{\text{dih}} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi, \theta\right].$$

Optimal values are defined as $V^*_{\phi} = \sup_{\pi \in \Pi} \inf_{\theta \in \Theta} V^{\pi,\theta}_{\phi}$. To the best of our knowledge, it is as of yet unknown whether such optimal values and their policies exist for every RPOMDP. Various RPOMDP papers claim the existence of an optimal value for their specific RPOMDP, but these results do not extend to the general RPOMDPs we consider in this paper [Osogami, 2015, Nakao *et al.*, 2021]. We prove that the optimal value for finite horizon exists for general RPOMDPs in Theorem 3.

By changing the stickiness or order of play of an RPOMDP, the optimal value may change:

Theorem 1 (Uncertainty assumptions matter)**.** *For an RPOMDP M, let $V^{*,M}_{fh}$ denote its optimal value for the finite horizon. In general, RPOMDPs with different stickiness functions, including static and dynamic uncertainty, may lead to different optimal values. Furthermore, a different order of play may also lead to different optimal values. Formally:*

1. *There exist RPOMDPs M_1, M_2 that only differ in their stickiness functions, such that $V^{*,M_1}_{fh} \neq V^{*,M_2}_{fh}$,*
2. *There exist RPOMDPs M_1, M_2 that only differ in their order of play, such that $V^{*,M_1}_{fh} \neq V^{*,M_2}_{fh}$.*

We sketch the proof here, for details see Appendix D.

Proof sketch. We construct explicit RPOMDPs and show that the optimal values do not coincide. For the first point regarding stickiness, consider the finite horizon RPOMDP in Figure 2. For zero stickiness, the value is $65\frac{1}{2}$, with agent policy $\pi = \{$○○◌◌ $\mapsto \{\mathrm{a} \mapsto \frac{1}{3}, \mathrm{b} \mapsto \frac{2}{3}\},$ ○●◌◌ $\mapsto \{\mathrm{a} \mapsto \frac{2}{3}, \mathrm{b} \mapsto \frac{1}{3}\}\}$ and nature policy $\theta = \{$○○◌ $\mapsto \{p \mapsto \frac{83}{270}, q \mapsto \frac{1}{10}\},$ ○●◌ $\mapsto \{p \mapsto \_, q \mapsto \frac{1}{3}\}\}$. For full stickiness, the value is $66\frac{2}{3}$, with the same agent policy π but different nature policy $\theta = \{$○ $\mapsto \{p \mapsto \frac{1}{3}, q \mapsto \frac{1}{3}\}\}$.

For the order of play, consider the finite horizon RPOMDP in Figure 3. For the agent first, the value is 30, with nature policy $\theta = \{\langle$○$, \mathrm{a}\rangle \mapsto \{p \mapsto 0.1\}, \langle$○$, \mathrm{b}\rangle \mapsto \{p \mapsto 0.9\}\}$ and any agent policy. For nature first, the value is 150, with nature policy $\theta = \{$○ $\mapsto \{p \mapsto 0.5\}\}$ and agent policy $\pi = \{$○ $\mapsto \{\mathrm{a} \mapsto 0.5, \mathrm{b} \mapsto 0.5\}\}$. □

Note that the optimal nature policies in these two RPOMDPs are deterministic. We show in Appendix D that deterministic policies suffice in these specific RPOMDPs due to the linearity of the value function in the nature policies. Furthermore, the RPOMDP we use to show that the order of play matters is fully observable and non-rectangular. In Appendix D, we show that the order of play still matters under some form of rectangularity.

Remark 4. For (s,a)-rectangular RMDPs, [Iyengar, 2005, Theorem 2.2] shows that static and dynamic semantics in RMDPs lead to the same optimal value. Iyengar establishes that in (s,a)-rectangular RMDPs, memoryless policies are sufficient for the agent. In response, nature may also play memoryless, as there is no incentive for nature to change its choice after its initial choice. As a consequence, zero and full stickiness coincide. This statement does not apply to RPOMDPs, where agents generally use memory. As shown in the (s,a)-rectangular RPOMDP in Figure 2 and Theorem 1, the optimal nature policy in this model's zero stickiness case uses information from previous observations, resulting in a smaller reward.

4 POSG Semantics for RPOMDPs

We formalize the underlying game of an RPOMDP as a zero-sum two-sided partially observable stochastic game (POSG) [Delage *et al.*, 2023], which is more widely studied than RPOMDPs. Our transformation allows us to carry over results from POSGs to RPOMDPs. In particular, we prove that our POSGs always have a Nash equilibrium for the finite horizon objective, which shows that optimal values and agent policies always exist in our finite horizon RPOMDPs.

Tracking fixed assignments. In our game, we explicitly keep track of the fixed variable assignments $u^{\hookleftarrow}$. The update function $\mathsf{upd}\colon \boldsymbol{U}^{\hookleftarrow} \times \boldsymbol{U} \times Z^{\mathfrak{n}}_{\bullet} \times Z_{\circ} \times A \to \boldsymbol{U}^{\hookleftarrow}$ updates the restricted variables after each valid nature choice following the stickiness of the RPOMDP M.

$$\mathsf{upd}(u^{\hookleftarrow}, u, z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a)(v) = \begin{cases} u(v) & \text{if } v \in U_h^{\text{stick}}(z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a), \\ u^{\hookleftarrow}(v) & \text{otherwise.} \end{cases}$$

By construction, recursively applying the update function on a path τ yields $\mathsf{fix}(\tau)$.

Definition 5. *Given an (agent first) RPOMDP* $\langle S, A, \boldsymbol{T}, R, Z^{\mathfrak{a}}_{\bullet}, Z^{\mathfrak{n}}_{\bullet}, Z_{\circ}, O^{\mathfrak{a}}_{\bullet}, O^{\mathfrak{n}}_{\bullet}, O_{\circ}, \mathsf{stick}, \mathfrak{a}\rangle$, *we define the POSG* $\langle \mathcal{S}^{\mathfrak{a}}, \mathcal{S}^{\mathfrak{n}}, \mathcal{A}^{\mathfrak{a}}, \mathcal{A}^{\mathfrak{n}}, \mathcal{T}, \mathcal{R}, \mathcal{Z}^{\mathfrak{a}}, \mathcal{Z}^{\mathfrak{n}}, \mathcal{O}^{\mathfrak{a}}, \mathcal{O}^{\mathfrak{n}}\rangle$, *with a set* $\mathcal{S}^{\mathfrak{a}} = S \times \boldsymbol{U}^{\hookleftarrow}$ *of* agent states, *a set* $\mathcal{S}^{\mathfrak{n}} = S \times \boldsymbol{U}^{\hookleftarrow} \times A$ *of* nature states, *a finite set* $\mathcal{A}^{\mathfrak{a}} = A$ *of* agent actions, *and a set* $\mathcal{A}^{\mathfrak{n}} = \boldsymbol{U}$ *of* nature actions. *The observations are defined as follows:* $\mathcal{Z}^{\mathfrak{a}} = Z^{\mathfrak{a}}_{\bullet} \times Z_{\circ}$ *is the finite set of the agent's observations, and* $\mathcal{Z}^{\mathfrak{n}} = Z^{\mathfrak{n}}_{\bullet} \times Z_{\circ} \times (A \cup \bot)$ *the finite set of nature's observations. The transition, reward, and observation functions are then defined as:*

- $\mathcal{T} = \mathcal{T}^{\mathfrak{a}} \cup \mathcal{T}^{\mathfrak{n}}$, *the transition function, where* $\mathcal{T}^{\mathfrak{a}}\colon \mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \to \mathcal{S}^{\mathfrak{n}}$ *is the agent's transition function, defined by* $\mathcal{T}^{\mathfrak{a}}(\langle s, u^{\hookleftarrow}\rangle, a) = \langle s, u^{\hookleftarrow}, a\rangle \in S^{\mathfrak{n}}$ *and* $\mathcal{T}^{\mathfrak{n}}\colon \mathcal{S}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}} \to \Delta(\mathcal{S}^{\mathfrak{a}})$ *is nature's transition function, such that* $\mathcal{T}^{\mathfrak{n}}(\langle s, u^{\hookleftarrow}, a\rangle, u, \langle s', \mathsf{upd}(u^{\hookleftarrow}, u, O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a)\rangle) =$
$$\begin{cases} \boldsymbol{T}(u)(s,a,s') & \text{if } u \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookleftarrow}), \\ 0 & \text{otherwise.} \end{cases}$$
- $\mathcal{R}\colon \mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \to \mathbb{R}$ *the reward function, given by* $\mathcal{R}(\langle s, u^{\hookleftarrow}\rangle, a) = R(s,a)$. *State-action pairs* $S^{\mathfrak{n}} \times A^{\mathfrak{n}}$ *have zero reward.*
- $\mathcal{O}^{\mathfrak{a}}\colon (\mathcal{S}^{\mathfrak{a}} \cup \mathcal{S}^{\mathfrak{n}}) \to \mathcal{Z}^{\mathfrak{a}}$ *the deterministic observations function of the agent defined as:*
$$\mathcal{O}^{\mathfrak{a}}(s) = \begin{cases} \langle O^{\mathfrak{a}}_{\bullet}(s'), O_{\circ}(s')\rangle & \text{if } s = \langle s', u^{\hookleftarrow}\rangle \in \mathcal{S}^{\mathfrak{a}}, \\ \langle O^{\mathfrak{a}}_{\bullet}(s'), O_{\circ}(s')\rangle & \text{if } s = \langle s', u^{\hookleftarrow}, a\rangle \in \mathcal{S}^{\mathfrak{n}}. \end{cases}$$
- $\mathcal{O}^{\mathfrak{n}}\colon (\mathcal{S}^{\mathfrak{a}} \cup \mathcal{S}^{\mathfrak{n}}) \to \mathcal{Z}^{\mathfrak{n}}$ *the deterministic observations function of nature defined as:*
$$\mathcal{O}^{\mathfrak{n}}(s) = \begin{cases} \langle O^{\mathfrak{n}}_{\bullet}(s'), O_{\circ}(s'), \bot\rangle & \text{if } s = \langle s', u^{\hookleftarrow}\rangle \in \mathcal{S}^{\mathfrak{a}}, \\ \langle O^{\mathfrak{n}}_{\bullet}(s'), O_{\circ}(s'), a\rangle & \text{if } s = \langle s', u^{\hookleftarrow}, a\rangle \in \mathcal{S}^{\mathfrak{n}}. \end{cases}$$

Game behavior. This game starts in an $\mathcal{S}^{\mathfrak{a}}$ state consisting of the initial state $s_I \in S$ of the RPOMDP and the totally undefined variable assignment $u^{\bot} \in \boldsymbol{U}^{\hookleftarrow}$. At any agent state $\langle s, u^{\hookleftarrow}\rangle$, both players observe their private and public observations of state s. After the agent chooses their action a, the game transitions deterministically to a nature state $\langle s, u^{\hookleftarrow}, a\rangle$. Again, both players observe their private and public observations of state s, while nature observes the agent's last action. Next, nature selects a variable assignment $u \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookleftarrow})$ from the set of variable assignments that agree with nature's past choices and hence account for the stickiness of the RPOMDP. Then the uncertain transition function $\boldsymbol{T}$ is resolved with u after which the game stochastically moves to the next agent state $\langle s', \mathsf{upd}(u^{\hookleftarrow}, u, O^{\mathfrak{n}}_{\bullet}(s), O^{\circ}(s), a)\rangle$, where s' can be reached from s given action a and the resolved transition function.

Nature chooses first. The POSG above is defined for RPOMDPs where the agent plays first, where $\mathsf{play} = \mathfrak{a}$. As nature can observe the agent's action choice, it may use this information to choose a transition function from the uncertainty set. If we assume that nature plays first, this information is not available yet; hence, the structure of the POSG needs to be changed to reflect this. For the remainder of the main paper, we focus on the case where the agent moves first, *i.e.*, RPOMDPs with $\mathsf{play} = \mathfrak{a}$. Our results carry over to RPOMDPs, where nature moves first. See Appendix E.

Paths and histories. A path in a POSG is a sequence of successive states and actions that alternate between agent and nature: $\langle s^{\mathfrak{a}}_0, a^{\mathfrak{a}}_0, s^{\mathfrak{n}}_0, a^{\mathfrak{n}}_0, s^{\mathfrak{a}}_1, a^{\mathfrak{a}}_1, \dots\rangle \in (\mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \times \mathcal{S}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}})^* \times \mathcal{S}^{\mathfrak{a}}$. A path is valid if $\forall s^{\mathfrak{a}}_i, a^{\mathfrak{a}}_i, s^{\mathfrak{n}}_i.\, \mathcal{T}^{\mathfrak{a}}(s^{\mathfrak{a}}_i, a^{\mathfrak{a}}_i, s^{\mathfrak{n}}_i) > 0$, and $\forall s^{\mathfrak{n}}_i, a^{\mathfrak{n}}_i, s^{\mathfrak{a}}_{i+1}.\, \mathcal{T}^{\mathfrak{n}}(s^{\mathfrak{n}}_i, a^{\mathfrak{n}}_i, s^{\mathfrak{a}}_{i+1}) > 0$. The set of paths in G is Paths^G. In the POSGs we consider, players only observe their own actions. A history for the agent or nature is a path mapped to their respective observations: the agent only observes agent actions, and their histories are sequences of the form $\langle \mathcal{O}^{\mathfrak{a}}(s^{\mathfrak{a}}_0), a^{\mathfrak{a}}_0, \mathcal{O}^{\mathfrak{a}}(s^{\mathfrak{n}}_0), \mathcal{O}^{\mathfrak{a}}(s^{\mathfrak{a}}_1), a^{\mathfrak{a}}_1, \mathcal{O}^{\mathfrak{a}}(s^{\mathfrak{n}}_1), \dots\rangle \in (\mathcal{Z}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{a}})^* \times \mathcal{Z}^{\mathfrak{a}}$, while the histories of nature are sequences in $(\mathcal{Z}^{\mathfrak{n}} \times \mathcal{Z}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}})^* \times \mathcal{Z}^{\mathfrak{n}}$. The sets of agent and nature histories in G are $H^{\mathfrak{a},G}$ and $H^{\mathfrak{n},G}$, respectively.

Policies and values. A policy for the agent in POSG G is a function $\pi\colon H^{\mathfrak{a},G} \to \Delta(A^{\mathfrak{a}})$, and a policy for nature is a function $\theta\colon H^{\mathfrak{n},G} \to \Delta(A^{\mathfrak{n}})$. The sets of all agent and nature policies in G are denoted by Π^G and Θ^G, respectively. The sets of deterministic and mixed policies are constructed analogously as for POMDPs in Section 2. The value of a

POSG is the expected reward collected under both players' policies π, θ:

$$V_{\text{fh}}^{\pi,\theta} = \mathbb{E}\left[\sum_{t=0}^{K-1} r_t \mid \pi, \theta\right], \quad V_{\text{dih}}^{\pi,\theta} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi, \theta\right].$$

4.1 Correctness of the Transformation

In the following, we show the correctness of our transformation from RPOMDP to POSG. We do this by (1) constructing bijections between paths and histories of an RPOMDP and its POSG, (2) using these bijections to derive bijections between the agent and nature policies for both RPOMDP and POSG, and (3) concluding with an equivalence between the values for both models. All proofs, including the explicit construction of all bijections, can be found in Appendix F.

Proposition 1 (Bijection between paths and histories)**.** *Let M be an RPOMDP, and G the POSG of M. There exists a bijection $f\colon Paths^M \to Paths^G$ and bijections between individual players' histories:*

- *Let $H^{\mathfrak{a},M}$ and $H^{\mathfrak{a},G}$ be the set of all agent histories in M and G, respectively. There exists a bijection $f^{\mathfrak{a},h}\colon H^{\mathfrak{a},M} \to H^{\mathfrak{a},G}$.*
- *Let $H^{\mathfrak{n},M}$ and $H^{\mathfrak{n},G}$ be the set of all nature histories in M and G, respectively. There exists a bijection $f^{\mathfrak{n},h}\colon H^{\mathfrak{n},M} \to H^{\mathfrak{n},G}$.*

Using the bijection between histories, we relate agent policies Π^M with Π^G and nature policies Θ^M with Θ^G.

Proposition 2 (Bijection between policies)**.** *Let M be an RPOMDP, and G the POSG of M. There exist bijections $f^\pi\colon \Pi^M \to \Pi^G$ and $f^\theta\colon \Theta^M \to \Theta^G$ between the agent's and nature's policies in M and G, respectively.*

An agent RPOMDP policy $\pi^M \in \Pi^M$ and an agent POSG policy $\pi^G \in \Pi^G$ are *corresponding* if π^M maps to π^G via the bijection f^π, *i.e.*, $\pi^G = f^\pi(\pi^M)$. Similarly, a nature RPOMDP policy θ^M and a nature POSG policy θ^G are *corresponding* if $\theta^G = f^\theta(\theta^M)$. From Proposition 2 it then follows that for two corresponding agent policies and two corresponding nature policies, the values of the RPOMDP and the POSG coincide.

Theorem 2 (Equivalent values)**.** *Let M be an RPOMDP, and G the POSG of M. Let $\pi^M \in \Pi^M, \pi^G = f^\pi(\pi^M) \in \Pi^G$ be corresponding agent policies, and $\theta^M \in \Theta^M, \theta^G = f^\theta(\theta^M) \in \Theta^G$ be corresponding nature policies. Then, their values for the RPOMDP and POSG coincide:*

$$V_\phi^{\pi^M,\theta^M} = V_\phi^{\pi^G,\theta^G}.$$

By showing that there is a bijection between RPOMDP and POSG policies and that the values coincide, we have established that these POSGs form an operational model for RPOMDP semantics.

4.2 Existence of Nash Equilibria

Using the RPOMDP to POSG transformation, we prove the existence of optimal values and policies for the agent in an RPOMDP for the finite horizon objective. That is, the existence of maximal values agent policies can achieve against all nature policies, such that $V_{\text{fh}}^* = \sup_{\pi\in\Pi} \inf_{\theta\in\Theta} \mathbb{E}[\sum_{t=0}^{K-1} r_t \mid \pi, \theta]$. From Theorem 2, it follows that if the values V_ϕ^* exist in the POSG G of an RPOMDP M, they also exist in M.

The value $V_\phi^{\pi,\theta}$ of a POSG G is a *Nash equilibrium*, and both players' policies are Nash optimal, denoted π^*, θ^*, if there is no incentive for either player to change their policy. That is, for either objective $\phi \in \{\text{fh}, \text{dih}\}$ we have:

$$\forall \pi \in \Pi^G . V_\phi^{\pi^*,\theta^*} \geq V_\phi^{\pi,\theta^*} \;\wedge\; \forall \theta \in \Theta^G . V_\phi^{\pi^*,\theta^*} \leq V_\phi^{\pi^*,\theta}.$$

Since the uncertainty set is infinite, our POSGs do not meet the standard requirements for a Nash equilibrium to exist [Peters, 2015, Fijalkow *et al.*, 2023]. Yet, our POSGs exhibit enough structure to show that a Nash equilibrium always exists for the finite horizon objective.

Theorem 3 (Existence of finite horizon Nash equilibrium)**.** *Let M be an RPOMDP and G the POSG of M. For the finite horizon objective $V_{\text{fh}}^{\pi,\theta} = \sum_{t=0}^{k-1}[r_t \mid \pi, \theta]$ we have the following saddle point condition in G:*

$$\sup_{\pi\in\Pi^G} \inf_{\theta\in\Theta^G} V_{\text{fh}}^{\pi,\theta} = \inf_{\theta\in\Theta^G} \sup_{\pi\in\Pi} V_{\text{fh}}^{\pi,\theta}. \tag{1}$$

From Equation (1)*, the existence of a Nash equilibrium in G follows immediately [Peters, 2015].*

We sketch the proof here; for details see Appendix G.

Proof sketch. We show the existence of the Nash equilibrium for our RPOMDPs by first defining a sufficient statistic (Appendix G.1). This statistic tracks histories and nature's policy and is an adaptation of the definition of [Delage *et al.*, 2023]. We use the sufficient statistic to construct the state space of a non-observable occupancy game (Appendix G.2) between agent and nature. Additionally, we show that we can reason about the optimal value and policies of the occupancy game, and thus those of the POSG, with the sets of mixed agent and nature policies instead of the sets of stochastic policies (Appendix G.3). Using the sets of mixed policies, we show that the constructed occupancy game is a semi-infinite convex game, as defined by [Lopez and Vercher, 1986] (Appendix G.4). Finally, we show that our occupancy game meets the conditions given by [Lopez and Vercher, 1986] for the existence of a saddle point. From the existence of the saddle point, the existence of the Nash equilibrium and an optimal agent policy immediately follows [Peters, 2015]. □

Whether a Nash equilibrium exists in the POSG G for discounted infinite horizon objective $V_{dih}^{\pi,\theta} = \sum_{t=0}^{\infty}[\gamma^t r_t \mid \pi, \theta]$ or a saddle point condition that would imply this Nash equilibrium remains an open problem.

Other semantic implications for RPOMDPs. To shed light on the reason why two RPOMDPs that only differ in either their stickiness or order of play can lead to different optimal values, we look at the structure of the POSGs of these RPOMDPs. Specifically, the RPOMDP from Figure 2 with either zero or full stickiness leads to the two POSGs depicted in Figure 4. The key difference between these POSGs is that in the zero stickiness case, every variable assignment by nature leads to the same two successor states, while in the full

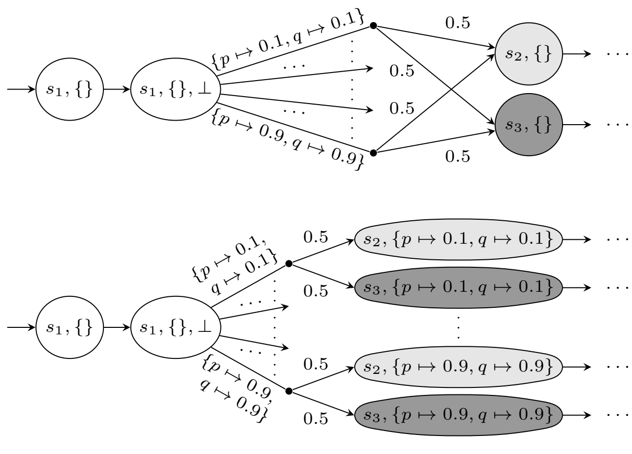

Figure 4: First states of zero stickiness (top) and full stickiness (bottom) POSGs of the RPOMDP in Figure 2.

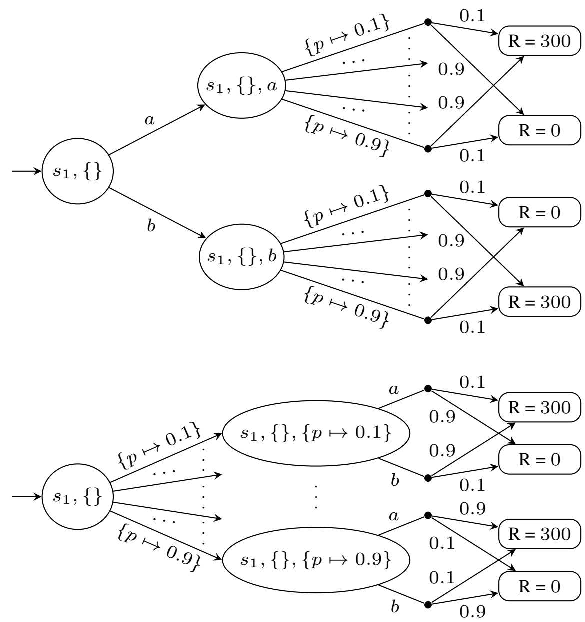

Figure 5: Agent first (top) and nature first (bottom) POSGs of the RPOMDP in Figure 3.

Reference	Stickiness	Order of play
[Osogami, 2015]	Zero	Agent first
[Chamie and Mostafa, 2018]	Zero	Agent first
[Saghafian, 2018]	Zero	Agent first
[Nakao *et al.*, 2021]	Zero	Agent first
[Suilen *et al.*, 2020]	Full	Nature first
[Cubuktepe *et al.*, 2021]	Full	Nature first

Table 1: Classification of existing RPOMDP literature.

stickiness case, any variable assignment by nature leads to two *unique* successor states and thus an infinitely branching POSG. A similar structural difference can be seen in the two POSGs depicted in Figure 5, which show the difference in the order of play for the RPOMDP in Figure 3.

5 Related Work

In this section, we first classify the existing RPOMDP literature into the different assumptions discussed in this paper, and then we provide a general overview of the related work.

5.1 Classification of RPOMDP Methods

Table 1 provides an overview of RPOMDP solution methods within our game semantics, specifically classifying the type of stickiness and the order of play for these methods.

Note that in the table, full stickiness and nature first order of play are always combined, as are zero stickiness and agent first order of play. This can be explained by those combinations being the most intuitive extensions of static and dynamic uncertainty to the partially observable setting. We also remark that [Saghafian, 2018] defines their problem by one fixed but unknown model that is chosen non-deterministically from the start, implying full stickiness and nature first, but their algorithmic solution method operates with zero stickiness and agent first semantics.

5.2 Further Related Work

The connection between rectangular RMDPs and stochastic games is well-established, see for instance [Iyengar, 2005, Xu and Mannor, 2010, Wiesemann *et al.*, 2013]. Yet, a key difference is that in RMDPs, nature is typically assumed to play stationary, as already mentioned in Section 2.2. This assumption is common because it is either sufficient for nature to play stationary or there are computational reasons. In SGs, on the other hand, history-based policies, as we also use, are common for both agent and nature. For a more elaborate discussion, see [Grand-Clément *et al.*, 2023, Section 2.2]. Recent work explores the connection between RMDPs and SGs in more depth [Chatterjee *et al.*, 2023].

For RPOMDPs, the connection with POSGs has also been alluded to before. [Osogami, 2015] briefly mention zero-sum games in their proof of convexity of the value function. [Saghafian, 2018] draws a link to nonzero-sum games, as they assume non-adversarial behavior for nature. [Rasouli and Saghafian, 2018] states a correspondence between the perfect Bayesian equilibrium in a zero-sum and the optimal value and policies in their RPOMDPs. Finally, [Nakao *et al.*, 2021] reasons about their RPOMDPs via games as well, but they assume the agent can observe nature's earlier choices.

6 Conclusion

This paper provides a semantic study of RPOMDPs, *i.e.*, the extension of RMDPs to the partially observable setting. We demonstrate that semantic choices that are irrelevant on RMDPs are important in RPOMDPs. We concretely provide semantics expressed as partially observable stochastic games and use this to derive novel results about the existence of Nash equilibria. Finally, we categorize algorithms from the literature based on our semantic framework. For future work, we aim to adapt solution methods for POSGs, like [Delage *et al.*, 2023], to solve RPOMDPs. We also plan to investigate the existence of a Nash equilibrium in the infinite horizon case.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback. This research has been partially funded by the NWO grant OCENW.KLEIN.187, the NWO Veni grant 222.147 (ProMiSe), and the ERC Starting Grant 101077178 (DEUCE).

References

[Behzadian *et al.*, 2021] Bahram Behzadian, Marek Petrik, and Chin Pang Ho. Fast algorithms for L_∞-constrained s-rectangular robust MDPs. In *NeurIPS*, pages 25982–25992, 2021.

[Bovy *et al.*, 2024] Eline M. Bovy, Marnix Suilen, Sebastian Junges, and Nils Jansen. Imprecise probabilities meet partial observability: Game semantics for robust POMDPs. *CoRR*, abs/2405.04941, 2024.

[Bovy, 2023] Eline M. Bovy. *The Underlying Belief Model of Uncertain Partially Observable Markov Decision Processes*. Master thesis, Radboud University, 2023.

[Chamie and Mostafa, 2018] Mahmoud El Chamie and Hala Mostafa. Robust action selection in partially observable Markov decision processes with model uncertainty. In *CDC*, pages 5586–5591. IEEE, 2018.

[Chatterjee *et al.*, 2016] Krishnendu Chatterjee, Martin Chmelik, Raghav Gupta, and Ayush Kanodia. Optimal cost almost-sure reachability in POMDPs. *Artif. Intell.*, 234:26–48, 2016.

[Chatterjee *et al.*, 2023] Krishnendu Chatterjee, Ehsan Kafshdar Goharshady, Mehrdad Karrabi, Petr Novotnỳ, and Đorđe Žikelić. Solving long-run average reward robust MDPs via stochastic games. *CoRR*, abs/2312.13912, 2023.

[Cubuktepe *et al.*, 2021] Murat Cubuktepe, Nils Jansen, Sebastian Junges, Ahmadreza Marandi, Marnix Suilen, and Ufuk Topcu. Robust finite-state controllers for uncertain POMDPs. In *AAAI*, pages 11792–11800. AAAI Press, 2021.

[Delage *et al.*, 2023] Aurélien Delage, Olivier Buffet, Jilles S. Dibangoye, and Abdallah Saffidine. HSVI can solve zero-sum partially observable stochastic games. *Dynamic Games and Applications*, 2023.

[Fijalkow *et al.*, 2023] Nathanaël Fijalkow, Nathalie Bertrand, Patricia Bouyer-Decitre, Romain Brenguier, Arnaud Carayol, John Fearnley, Hugo Gimbert, Florian Horn, Rasmus Ibsen-Jensen, Nicolas Markey, Benjamin Monmege, Petr Novotný, Mickael Randour, Ocan Sankur, Sylvain Schmitz, Olivier Serre, and Mateusz Skomra. Games on graphs. *CoRR*, abs/2305.10546, 2023.

[Gillette, 1957] Dean Gillette. Stochastic games with zero stop probabilities. *Contributions to the Theory of Games*, 3:179–187, 1957.

[Grand-Clément *et al.*, 2023] Julien Grand-Clément, Marek Petrik, and Nicolas Vieille. Beyond discounted returns: Robust Markov decision processes with average and blackwell optimality. *CoRR*, abs/2312.03618, 2023.

[Ho *et al.*, 2018] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Fast bellman updates for robust MDPs. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1984–1993. PMLR, 2018.

[Ho *et al.*, 2021] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for L_1-robust Markov decision processes. *J. Mach. Learn. Res.*, 22:275:1–275:46, 2021.

[Iyengar, 2005] Garud N. Iyengar. Robust dynamic programming. *Math. Oper. Res.*, 30(2):257–280, 2005.

[Jaksch *et al.*, 2010] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.

[Jansen *et al.*, 2022] Nils Jansen, Sebastian Junges, and Joost-Pieter Katoen. Parameter synthesis in Markov models: A gentle survey. In *Principles of Systems Design*, volume 13660 of *LNCS*, pages 407–437. Springer, 2022.

[Kaelbling *et al.*, 1998] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1-2):99–134, 1998.

[Kwiatkowska *et al.*, 2022] Marta Kwiatkowska, Gethin Norman, David Parker, Gabriel Santos, and Rui Yan. Probabilistic model checking for strategic equilibria-based decision making: Advances and challenges (invited talk). In *MFCS*, volume 241 of *LIPIcs*, pages 4:1–4:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[Lopez and Vercher, 1986] M.A. Lopez and D.E. Vercher. Convex semi-infinite games. *Journal of optimization theory and applications*, 50(2):289–312, 1986.

[Moos *et al.*, 2022] Janosch Moos, Kay Hansel, Hany Abdulsamad, Svenja Stark, Debora Clever, and Jan Peters. Robust reinforcement learning: A review of foundations and recent advances. *Mach. Learn. Knowl. Extr.*, 4(1):276–315, 2022.

[Nakao *et al.*, 2021] Hideaki Nakao, Ruiwei Jiang, and Siqian Shen. Distributionally robust partially observable Markov decision process with moment-based ambiguity. *SIAM J. Optim.*, 31(1):461–488, 2021.

[Nilim and Ghaoui, 2005] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.*, 53(5):780–798, 2005.

[Osogami, 2015] Takayuki Osogami. Robust partially observable Markov decision process. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 106–115. JMLR.org, 2015.

[Peters, 2015] Hans Peters. *Game Theory: A Multi-Leveled Approach*. Springer Texts in Business and Economics. Springer, second edition, 2015.

[Petrik and Subramanian, 2014] Marek Petrik and Dharmashankar Subramanian. RAAM: the benefits of robustness in approximating aggregated MDPs in reinforcement learning. In *NIPS*, pages 1979–1987, 2014.

[Puterman, 1994] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994.

[Rasouli and Saghafian, 2018] Mohammad Rasouli and Soroush Saghafian. Robust partially observable Markov decision processes. *HKS Working Paper*, RWP18-027, 2018.

[Saghafian, 2018] Soroush Saghafian. Ambiguous partially observable Markov decision processes: Structural results and applications. *J. Econ. Theory*, 178:1–35, 2018.

[Shapley, 1953] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[Spaan, 2012] Matthijs T. J. Spaan. Partially observable Markov decision processes. In *Reinforcement Learning*, volume 12 of *Adaptation, Learning, and Optimization*, pages 387–414. Springer, 2012.

[Suilen *et al.*, 2020] Marnix Suilen, Nils Jansen, Murat Cubuktepe, and Ufuk Topcu. Robust policy synthesis for uncertain POMDPs via convex optimization. In *IJCAI*, pages 4113–4120. ijcai.org, 2020.

[Suilen *et al.*, 2022] Marnix Suilen, Thiago D. Simão, David Parker, and Nils Jansen. Robust anytime learning of Markov decision processes. In *NeurIPS*, 2022.

[Wang *et al.*, 2023] Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust MDPs with global convergence guarantee. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 35763–35797. PMLR, 2023.

[Wiesemann *et al.*, 2013] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Math. Oper. Res.*, 38(1):153–183, 2013.

[Xu and Mannor, 2010] Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. In *NIPS*, pages 2505–2513. Curran Associates, Inc., 2010.