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 $\bot$ 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.

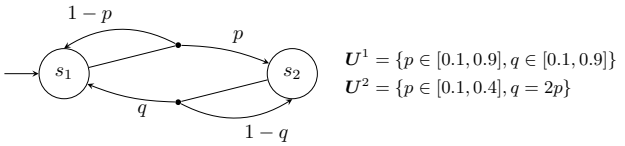


Figure 1: An example RMDP with two uncertainty sets.

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.

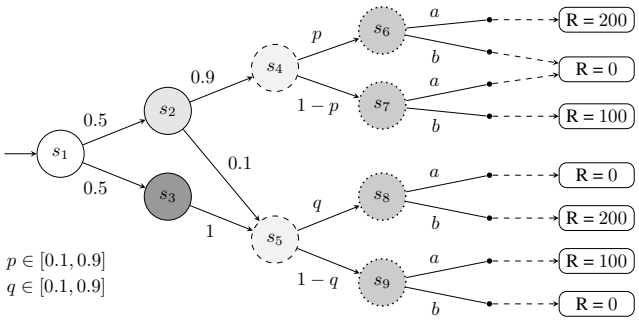


Figure 2: An RPOMDP where full and zero stickiness do not coincide in their optimal value.

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

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

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

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 $\tau = \langle s_0, a_0, u_0, s_1, \ldots, s_n \rangle \in \text{Paths}^M$. For $k < n$, we denote the prefix $\tau_{0:k} = \langle s_0, a_0, u_0, s_1, \ldots, s_k \rangle$. 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, $\forall h^{\mathfrak{n}} \in H^{\mathfrak{n}}, \forall a \in A, \forall u \in \boldsymbol{U}$.

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

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 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.

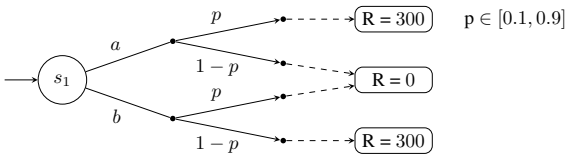


Figure 3: An RPOMDP where nature first and agent first semantics do not coincide in their optimal value.

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

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 $\pi = \{\bigcirc\bigcirc(\cdot)(\cdot) \mapsto \{a \mapsto \frac{1}{3}, b \mapsto \frac{2}{3}\}, \bigcirc\bullet(\cdot)(\cdot) \mapsto \{a \mapsto \frac{2}{3}, b \mapsto \frac{1}{3}\}\}$ and nature policy $\theta = \{\bigcirc\bigcirc(\cdot) \mapsto \{p \mapsto \frac{83}{270}, q \mapsto \frac{1}{10}\}, \bigcirc\bullet(\cdot) \mapsto \{p \mapsto \_, q \mapsto \frac{1}{3}\}\}$. For full stickiness, the value is $66\frac{2}{3}$, with agent policy $\pi = \{\bigcirc\bigcirc(\cdot)(\cdot) \mapsto \{a \mapsto \frac{1}{3}, b \mapsto \frac{2}{3}\}, \bigcirc\bullet(\cdot)(\cdot) \mapsto \{a \mapsto \frac{7}{10}, b \mapsto \frac{3}{10}\}$ and nature policy $\theta = \{\bigcirc \mapsto \{p \mapsto \frac{1}{3}, q \mapsto \frac{1}{3}\}\}$.

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

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 $u^{\hookrightarrow}$. The update function $\mathsf{upd}\colon \boldsymbol{U}^{\hookrightarrow} \times \boldsymbol{U} \times Z_\bullet^{\mathfrak{n}} \times Z_\circ \times A \to \boldsymbol{U}^{\hookrightarrow}$ updates the restricted variables after each valid nature choice following the stickiness of the RPOMDP M.

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

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

Definition 5. *Given an (agent first) RPOMDP $\langle S, A, \boldsymbol{T}, R, Z_\bullet^{\mathfrak{a}}, Z_\bullet^{\mathfrak{n}}, Z_\circ, O_\bullet^{\mathfrak{a}}, O_\bullet^{\mathfrak{n}}, O_\circ, \mathsf{stick}, \mathfrak{a}\rangle$, 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 $\mathcal{S}^{\mathfrak{a}} = S \times \boldsymbol{U}^{\hookrightarrow}$ of* agent states*, a set $\mathcal{S}^{\mathfrak{n}} = S \times \boldsymbol{U}^{\hookrightarrow} \times A$ 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: $\mathcal{Z}^{\mathfrak{a}} = Z_\bullet^{\mathfrak{a}} \times Z_\circ$ is the finite set of the agent's observations, and $\mathcal{Z}^{\mathfrak{n}} = Z_\bullet^{\mathfrak{n}} \times Z_\circ \times (A \cup \bot)$ 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 $\mathcal{T}^{\mathfrak{a}}(\langle s, u^{\hookrightarrow}\rangle, a) = \langle s, u^{\hookrightarrow}, a\rangle \in S^{\mathfrak{n}}$ 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 $\mathcal{T}^{\mathfrak{n}}(\langle s, u^{\hookrightarrow}, a\rangle, u, \langle s', \mathsf{upd}(u^{\hookrightarrow}, u, O_\bullet^{\mathfrak{n}}(s), O_\circ(s), a)\rangle) =$*
  $$\begin{cases} \boldsymbol{T}(u)(s,a,s') & \text{if } u \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookrightarrow}), \\ 0 & \text{otherwise.} \end{cases}$$
- $\mathcal{R}\colon \mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \to \mathbb{R}$ *the reward function, given by $\mathcal{R}(\langle s, u^{\hookrightarrow}\rangle, a) = R(s,a)$. 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:*
  $$\mathcal{O}^{\mathfrak{a}}(s) = \begin{cases} \langle O_\bullet^{\mathfrak{a}}(s'), O_\circ(s')\rangle & \text{if } s = \langle s', u^{\hookrightarrow}\rangle \in \mathcal{S}^{\mathfrak{a}}, \\ \langle O_\bullet^{\mathfrak{a}}(s'), O_\circ(s')\rangle & \text{if } s = \langle s', u^{\hookrightarrow}, a\rangle \in \mathcal{S}^{\mathfrak{n}}. \end{cases}$$
- $\mathcal{O}^{\mathfrak{n}}\colon (\mathcal{S}^{\mathfrak{a}} \cup \mathcal{S}^{\mathfrak{n}}) \to \mathcal{Z}^{\mathfrak{n}}$ *the deterministic observations function of nature defined as:*
  $$\mathcal{O}^{\mathfrak{n}}(s) = \begin{cases} \langle O_\bullet^{\mathfrak{n}}(s'), O_\circ(s'), \bot\rangle & \text{if } s = \langle s', u^{\hookrightarrow}\rangle \in \mathcal{S}^{\mathfrak{a}}, \\ \langle O_\bullet^{\mathfrak{n}}(s'), O_\circ(s'), a\rangle & \text{if } s = \langle s', u^{\hookrightarrow}, a\rangle \in \mathcal{S}^{\mathfrak{n}}. \end{cases}$$

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 $u^{\bot} \in \boldsymbol{U}^{\hookrightarrow}$. At any agent state $\langle s, u^{\hookrightarrow}\rangle$, 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 $\langle s, u^{\hookrightarrow}, a\rangle$. Again, both players observe their private and public observations of state s, with which nature observes the agent's last action. Next, nature selects a variable assignment $u \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookrightarrow})$ 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 $\langle s', \mathsf{upd}(u^{\hookrightarrow}, u, O_\bullet^{\mathfrak{n}}(s), O^\circ(s), a)\rangle$, 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: $\langle s_0^{\mathfrak{a}}, a_0^{\mathfrak{a}}, s_0^{\mathfrak{n}}, a_0^{\mathfrak{n}}, s_1^{\mathfrak{a}}, a_1^{\mathfrak{a}}, \ldots\rangle \in (\mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \times \mathcal{S}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}})^* \times \mathcal{S}^{\mathfrak{a}}$, A path is valid if $\forall s_i^{\mathfrak{a}}, a_i^{\mathfrak{a}}, s_i^{\mathfrak{n}}.\, \mathcal{T}^{\mathfrak{a}}(s_i^{\mathfrak{a}}, a_i^{\mathfrak{a}}, s_i^{\mathfrak{n}}) > 0$, and $\forall s_i^{\mathfrak{n}}, a_i^{\mathfrak{n}}, s_{i+1}^{\mathfrak{a}}.\, \mathcal{T}^{\mathfrak{n}}(s_i^{\mathfrak{n}}, a_i^{\mathfrak{n}}, s_{i+1}^{\mathfrak{a}}) > 0$. 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 $\langle \mathcal{O}^{\mathfrak{a}}(s_0^{\mathfrak{a}}), a_0^{\mathfrak{a}}, \mathcal{O}^{\mathfrak{a}}(s_0^{\mathfrak{n}}), \mathcal{O}^{\mathfrak{a}}(s_1^{\mathfrak{a}}), a_1^{\mathfrak{a}}, \mathcal{O}^{\mathfrak{a}}(s_1^{\mathfrak{n}}), \ldots\rangle \in (\mathcal{Z}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{a}})^* \times \mathcal{Z}^{\mathfrak{a}}$, 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 $\theta\colon H^{\mathfrak{n},G} \times \mathcal{Z}^{\mathfrak{n}} \to \Delta(A^{\mathfrak{n}})$. 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 $f\colon \mathit{Paths}^M \to \mathit{Paths}^G$ 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:*

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

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

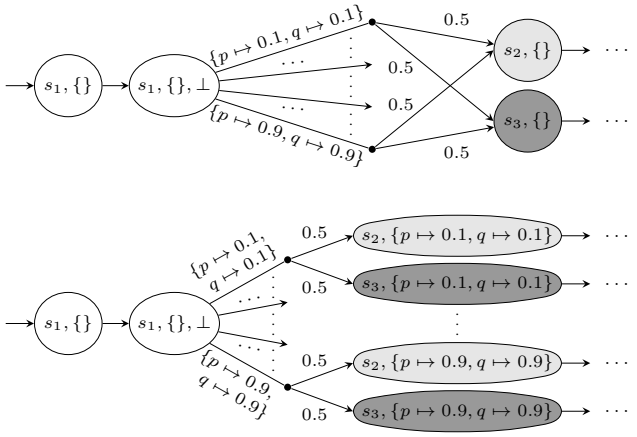


Figure 4: First states of zero stickiness (top) and full stickiness (bottom) POSGs of the RPOMDP in Figure 2.

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].

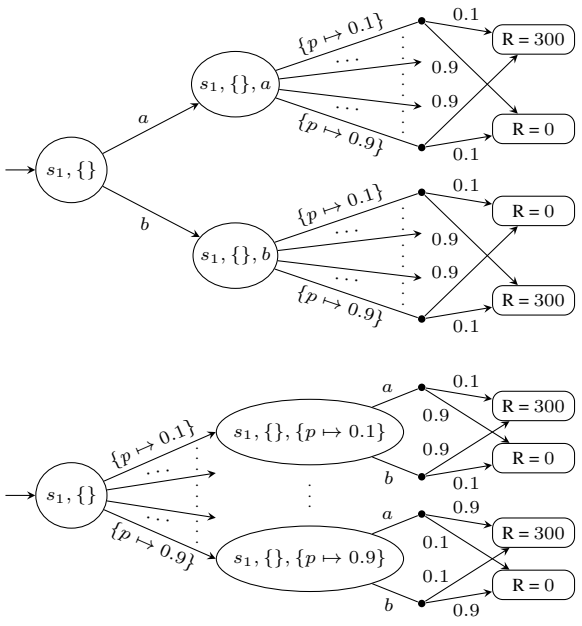


Figure 5: Agent first (top) and nature first (bottom) POSGs of the RPOMDP in Figure 3.

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.

[GeoGebra GmbH, 2024] GeoGebra GmbH. Geogebra (online), 2024. Available at https://www.geogebra.org.

[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.

[Kuhn, 1953] Harold W Kuhn. Extensive games and the problem of information. *Contributions to the Theory of Games*, 2(28):193–216, 1953.

[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.

[Zalinescu, 2002] Constantin Zalinescu. *Convex analysis in general vector spaces*. World scientific, 2002.

A Appendix Overview & Additional Preliminaries

These appendices contain additional results and proofs for the claims made in the main text. In particular, the appendices are structured as follows:

- Appendix A contains an overview of key notation used and additional preliminaries (mostly around paths and histories) used in subsequent appendices.
- Appendix B show how an arbitrary RPOMDP, with uncertainty in both transitions and observations, can be transformed into an RPOMDP with deterministic state-based observations as we use throughout the main text inline with Definition 3.
- Appendix C defines observation-based stickiness and shows its workings in an example.
- Appendix D details the result from Theorem 1. That is, it shows that different forms of stickiness or a different order of play can lead to different optimal values in RPOMDPs.
- Appendix E discussed the required changes to the constructions from the main paper when considering RPOMDPs with nature first semantics.
- Appendix F contains the proof of Theorem 2 and the propositions it builds on.
- Appendix G contains the proof of Theorem 3.

A.1 Glossary of Key Notation

Notation	Description
$K \in \mathbb{N}$	Finite horizon bound
γ	Discount factor
$\phi \in \{\text{fh,dih}\}$	Objective, either finite-horizon (fh) or discounted infinite horizon (dih)
$v \in U$	Variable v in the set of variables U
$\boldsymbol{U} \subseteq (U \to \mathbb{R})$	Uncertainty set, defined as a set of admissible variable assignments $u \in \boldsymbol{U}$
$\boldsymbol{T}(u) \colon S \times A \to \Delta(S)$	The uncertain transition function $\boldsymbol{T}$ instantiated by variable assignment u
$\boldsymbol{U}^{\hookrightarrow}$	Set of partial variable assignments $U \hookrightarrow \mathbb{R}$
$u^{\perp}$	The completely undefined variable assignment
$\boldsymbol{U}^{\mathcal{P}}(u^{\hookrightarrow})$	Set of variable assignments that agree with partial assignment $u^{\hookrightarrow}$
$\lvert\tau\rvert$	Horizon length of a path τ
$\lvert h\rvert$	Horizon length of a history h
$\oplus$	Concatenation of two tuples or paths
$\tau_{0:k}$	Prefix of horizon length k of path τ
H_t	The subset of histories of length t
$H_{0:t}$	The subset of histories of length 0 to t
$\pi \in \Pi$	Stochastic agent policies
$\theta \in \Theta$	Stochastic nature policies
$\pi^{det} \in \Pi^{det}$	Deterministic agent policies
$\theta^{det} \in \Theta^{det}$	Deterministic nature policies
$\pi^{mix} \in \Pi^{mix}$	Mixed agent policies
$\theta^{mix} \in \Theta^{mix}$	Mixed nature policies
π_t, θ_t	A policy defined on histories of length t (can be part of a larger policy π/θ)
$\pi_{0:t}, \theta_{0:t}$	A policy defined on histories of length 0 to t (can be part of a larger policy π/θ)
Π_t, Θ_t	The set of policies defined on histories of length t
$\Pi_{0:t}, \Theta_{0:t}$	The set of policies defined on histories of length 0 to t
$\mathfrak{a}$, $\mathfrak{n}$	Agent and nature
$z_\bullet$, $z_\circ$	Private and public observations
$\vec{x} = \langle x_0, x_1, \ldots, x_{n-1} \rangle \in \mathbb{R}^n$	Vector

Table 2: Glossary of key notation.

A.2 Additional Preliminaries

This appendix contains additional definitions and concepts used throughout the rest of the appendices.

Rectangularity. Rectangularity concerns dependencies between transitions in the model. If all transitions originating from different states are independent, we call the model and the uncertainty set s-rectangular. An s-rectangular uncertainty set $\boldsymbol{U}$ can be rewritten as the Cartesian product of smaller, s related uncertainty sets:

$$\boldsymbol{U} = \bigtimes_{s \in S} \boldsymbol{U}^s.$$

Similarly, if all transitions following different actions are independent, we call the model and the uncertainty set a-rectangular [Wiesemann *et al.*, 2013]. An a-rectangular uncertainty set $\boldsymbol{U}$ can be rewritten as the Cartesian product of smaller, a related uncertainty sets:

$$\boldsymbol{U} = \bigtimes_{a \in A} \boldsymbol{U}^a.$$

Combining these two, so if all transitions originating from different states and following from different actions are independent, we call the model and the uncertainty set s, a-rectangular [Iyengar, 2005]. An s, a-rectangular uncertainty set $\boldsymbol{U}$ can be rewritten as the Cartesian product of smaller, s and a related uncertainty sets:

$$\boldsymbol{U} = \bigtimes_{\substack{s \in S, \\ a \in A}} \boldsymbol{U}^{s,a}.$$

If we do not have any known independencies, we call the model and the uncertainty set non-rectangular.

Belief. A belief is a distribution over the set of states $\Delta(S)$ representing the probability of being in a state given the history. We compute a belief given an history by repeatedly applying the belief update [Kaelbling *et al.*, 1998], starting from the initial belief based on the initial state. Note that we adjusted the belief update to work with three observation functions.

Definition 6 (Belief update)**.** *Given a belief b, an agent action a, a nature action u, and observations $z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}$, we compute the successor belief b' for each state $s' \in S$ as follows:*

$$\begin{aligned} b'(s') &= Pr(\, s' \mid b, a, u, z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}\,) \\ &= \frac{O^{\mathfrak{a}}_{\bullet}(s', z^{\mathfrak{a}}_{\bullet}) O^{\mathfrak{n}}_{\bullet}(s', z^{\mathfrak{n}}_{\bullet}) O_{\circ}(s', z_{\circ}) \sum_{s \in S} b(s) \cdot \boldsymbol{T}(u)(s, a, s')}{\sum_{s' \in S} O^{\mathfrak{a}}_{\bullet}(s', z^{\mathfrak{a}}_{\bullet}) O^{\mathfrak{n}}_{\bullet}(s', z^{\mathfrak{n}}_{\bullet}) O_{\circ}(s', z_{\circ}) \sum_{s \in S} b(s) \cdot \boldsymbol{T}(u)(s, a, s')}. \end{aligned}$$

We use the belief update in our proof of the existence of a Nash equilibrium in Appendix G.

Dirac distribution. A distribution is called *Dirac* if it assigns probability one to precisely one element. We use Dirac distributions in our proof of the existence of a Nash equilibrium in Appendix G.

Graph-preserving Given an RPOMDP $M = \langle S, A, \boldsymbol{T}, R, Z^{\mathfrak{a}}_{\bullet}, Z^{\mathfrak{n}}_{\bullet}, Z_{\circ}, O^{\mathfrak{a}}_{\bullet}, O^{\mathfrak{n}}_{\bullet}, O_{\circ}, \mathsf{stick}, \mathsf{play} \rangle$ with uncertainty set $\boldsymbol{U}$, the uncertainty set is called graph-preserving if all variable assignments preserve the underlying structure of the RPOMDP. In other words, if a transition is possible given one variable assignment, it is possible given all variable assignment:

$$\forall s, s' \in S, \forall a \in A. \, (\exists u \in \boldsymbol{U}. \, \boldsymbol{T}(u)(s, a, s') = 0 \implies \forall u \in \boldsymbol{U}. \, \boldsymbol{T}(u)(s, a, s') = 0)\,.$$

Convex set A subset X of a Euclidean space is convex if given any two elements in the set the line drawn between those two elements is entirely contained in the set. Given 2 elements $x_0, x_1 \in X$ and scalar $\alpha \in [0, 1]$ we have that:

$$\alpha x_0 + (1 - \alpha) x_1 \in X.$$

As a result, a convex set has the property that any convex combination of its elements is again contained in that set. Given k elements $x_0, x_1, \ldots, x_{k-1} \in X$ and k non-negative scalars $\lambda_0, \lambda_1, \ldots, \lambda_{k-1} \in [0, 1]$ such that $\sum_{i=0}^{k-1} \lambda_i = 1$, we have that:

$$\sum_{i=0}^{k} \lambda_i x_i \in X.$$

Joint histories. The *joint history* combines the agent and nature histories in a single sequence:

$$\begin{aligned} H^M &\subseteq (Z^{\mathfrak{a}}_{\bullet} \times Z^{\mathfrak{n}}_{\bullet} \times Z_{\circ} \times A \times \boldsymbol{U})^* \times Z^{\mathfrak{a}}_{\bullet} \times Z^{\mathfrak{n}}_{\bullet} \times Z_{\circ}, \\ H^G &\subseteq (\mathcal{Z}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}})^* \times \mathcal{Z}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{n}}. \end{aligned}$$

Neither player can observe the joint history. Given a joint history $h \in H^M$ of $h \in H^G$, we use the superscripts $\mathfrak{a}$ and $\mathfrak{n}$ to indicate the agent and nature observable parts of the history h. So we get $h^{\mathfrak{a}} \in H^{\mathfrak{a},M}$ (or $\in H^{\mathfrak{a},G}$) and $h^{\mathfrak{n}} \in H^{\mathfrak{n},M}$ (or $\in H^{\mathfrak{n},G}$). We use joint histories in Appendices F and G.

Paths to histories. The following six functions map paths to histories in the RPOMDP and POSGs. The sets of histories in the RPOMDP and POSGs are constructed by applying these mappings to the sets of paths. $\text{Paths}^{\ltimes}$ indicates the set of all path segments, see Appendix F.

Definition 7 (Paths to joint histories in RPOMDPs)**.**
Let $O^{M,\ltimes} : \mathit{Paths}^{M,\ltimes} \to H^{M,\ltimes}$ *defined by:*

$$\begin{aligned} O^{M,\ltimes}(\langle s \rangle) &= \langle O^{\mathfrak{a}}_{\bullet}(s), O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s) \rangle. \\ O^{M,\ltimes}(\langle s, a, u \rangle) &= \langle O^{\mathfrak{a}}_{\bullet}(s), O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a, u \rangle. \\ O^{M,\ltimes}(\langle s, a, u \rangle \oplus \tau^{M\prime}) &= O^{M,\ltimes}(\langle s, a, u \rangle) \oplus O^{M,\ltimes}(\tau^{M\prime}). \end{aligned}$$

Let $O^{M} : \mathit{Paths}^{M} \to H^{M}$ *defined by:*

$$\begin{aligned} O^{M}(\langle s \rangle) &= \langle O^{\mathfrak{a}}_{\bullet}(s), O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s) \rangle. \\ O^{M}(\langle s, a, u \rangle) &= \langle O^{\mathfrak{a}}_{\bullet}(s), O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a, u \rangle. \\ O^{M}(\langle s, a, u \rangle \oplus \tau^{M\prime}) &= O^{M}(\langle s, a, u \rangle) \oplus O^{M,\ltimes}(\tau^{M\prime}). \end{aligned}$$

Definition 8 (Paths to agent histories in RPOMDPs)**.**
Let $O^{\mathfrak{a},M,\ltimes} : \mathit{Paths}^{M,\ltimes} \to H^{\mathfrak{a},M,\ltimes}$ *defined by:*

$$\begin{aligned} O^{\mathfrak{a},M,\ltimes}(\langle s \rangle) &= \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s) \rangle. \\ O^{\mathfrak{a},M,\ltimes}(\langle s, a, u \rangle) &= \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s), a \rangle. \\ O^{\mathfrak{a},M,\ltimes}(\langle s, a, u \rangle \oplus \tau^{M\prime}) &= O^{\mathfrak{a},M,\ltimes}(\langle s, a, u \rangle) \oplus O^{\mathfrak{a},M,\ltimes}(\tau^{M\prime}). \end{aligned}$$

Let $O^{\mathfrak{a},M} : \mathit{Paths}^{M} \to H^{\mathfrak{a},M}$ *defined by:*

$$\begin{aligned} O^{\mathfrak{a},M}(\langle s \rangle) &= \langle O^{\mathfrak{a}}_{\bullet}(s), O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s) \rangle. \\ O^{\mathfrak{a},M}(\langle s, a, u \rangle) &= \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s), a \rangle. \\ O^{\mathfrak{a},M}(\langle s, a, u \rangle \oplus \tau^{M\prime}) &= O^{\mathfrak{a},M}(\langle s, a, u \rangle) \oplus O^{\mathfrak{a},M,\ltimes}(\tau^{M\prime}). \end{aligned}$$

Definition 9 (Paths to nature histories in RPOMDPs)**.**
Let $O^{\mathfrak{n},M,\ltimes} : \mathit{Paths}^{M,\ltimes} \to H^{\mathfrak{n},M,\ltimes}$ *defined by:*

$$\begin{aligned} O^{\mathfrak{n},M,\ltimes}(\langle s \rangle) &= \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s) \rangle. \\ O^{\mathfrak{n},M,\ltimes}(\langle s, a, u \rangle) &= \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a, u \rangle. \\ O^{\mathfrak{n},M,\ltimes}(\langle s, a, u \rangle \oplus \tau^{M\prime}) &= O^{\mathfrak{n},M,\ltimes}(\langle s, a, u \rangle) \oplus O^{M,\ltimes}(\tau^{M\prime}). \end{aligned}$$

Let $O^{\mathfrak{n},M} : \mathit{Paths}^{M} \to H^{\mathfrak{n},M}$ *defined by:*

$$\begin{aligned} O^{\mathfrak{n},M}(\langle s \rangle) &= \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s) \rangle. \\ O^{\mathfrak{n},M}(\langle s, a, u \rangle) &= \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a, u \rangle. \\ O^{\mathfrak{n},M}(\langle s, a, u \rangle \oplus \tau^{M\prime}) &= O^{\mathfrak{n},M}(\langle s, a, u \rangle) \oplus O^{\mathfrak{n},M,\ltimes}(\tau^{M\prime}). \end{aligned}$$

Definition 10 (Paths to joint histories in POSGs)**.**
Similarly, let $O^{G,\ltimes} : \mathit{Paths}^{G,\ltimes} \to H^{G,\ltimes}$ *defined by:*

$$\begin{aligned} O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow} \rangle\rangle) &= \langle \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s) \rangle, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot \rangle \rangle. \\ O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow} \rangle, a, \langle s, u^{\hookrightarrow}, a \rangle, u \rangle) &= \langle \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s) \rangle, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot \rangle, a, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s) \rangle, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a \rangle, u \rangle. \\ O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow} \rangle, a, \langle s, u^{\hookrightarrow}, a \rangle, u \rangle \oplus \tau^{G\prime}) &= O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow} \rangle, a, \langle s, u^{\hookrightarrow}, a \rangle, u \rangle) \oplus O^{G,\ltimes}(\tau^{G\prime}). \end{aligned}$$

Let $O^{G} : \mathit{Paths}^{G} \to H^{G}$ *defined by:*

$$\begin{aligned} O^{G}(\langle\langle s, u^{\hookrightarrow} \rangle\rangle) &= \langle \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s) \rangle, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot \rangle \rangle. \\ O^{G}(\langle\langle s, u^{\hookrightarrow} \rangle, a, \langle s, u^{\hookrightarrow}, a \rangle, u \rangle) &= \langle \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s) \rangle, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot \rangle, a, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s) \rangle, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a \rangle, u \rangle. \\ O^{G}(\langle\langle s, u^{\hookrightarrow} \rangle, a, \langle s, u^{\hookrightarrow}, a \rangle, u \rangle \oplus \tau^{G\prime}) &= O^{G}(\langle\langle s, u^{\hookrightarrow} \rangle, a, \langle s, u^{\hookrightarrow}, a \rangle, u \rangle) \oplus O^{G,\ltimes}(\tau^{G\prime}). \end{aligned}$$

Definition 11 (Paths to agent histories in POSGs)**.**
Similarly, let $O^{\mathfrak{a},G,\ltimes} : \mathit{Paths}^{G,\ltimes} \to H^{\mathfrak{a},G,\ltimes}$ *defined by:*

$$O^{\mathfrak{a},G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle\rangle) = \langle\langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle.$$
$$O^{\mathfrak{a},G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle) = \langle\langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle, a, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle.$$
$$O^{\mathfrak{a},G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle \oplus \tau^{G'}) = O^{\mathfrak{a},G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle) \oplus O^{\mathfrak{a},G,\ltimes}(\tau^{G'}).$$

Let $O^{\mathfrak{a},G} : \mathit{Paths}^{G} \to H^{\mathfrak{a},G}$ *defined by:*

$$O^{\mathfrak{a},G}(\langle\langle s, u^{\hookrightarrow}\rangle\rangle) = \langle\langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle.$$
$$O^{\mathfrak{a},G}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle) = \langle\langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle, a, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle.$$
$$O^{\mathfrak{a},G}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle \oplus \tau^{G'}) = O^{\mathfrak{a},G}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle) \oplus O^{\mathfrak{a},G,\ltimes}(\tau^{G'}).$$

Definition 12 (Paths to nature histories in POSGs)**.**
Similarly, let $O^{\mathfrak{n},G,\ltimes} : \mathit{Paths}^{G,\ltimes} \to H^{\mathfrak{n},G,\ltimes}$ *defined by:*

$$O^{\mathfrak{n},G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle\rangle) = \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle\rangle.$$
$$O^{\mathfrak{n},G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle) = \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a\rangle, u\rangle.$$
$$O^{\mathfrak{n},G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle \oplus \tau^{G'}) = O^{\mathfrak{n},G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle) \oplus O^{\mathfrak{n},G,\ltimes}(\tau^{G'}).$$

Let $O^{\mathfrak{n},G} : \mathit{Paths}^{G} \to H^{\mathfrak{n},G}$ *defined by:*

$$O^{\mathfrak{n},G}(\langle\langle s, u^{\hookrightarrow}\rangle\rangle) = \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle\rangle.$$
$$O^{\mathfrak{n},G}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle) = \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a\rangle, u\rangle.$$
$$O^{\mathfrak{n},G}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle \oplus \tau^{G'}) = O^{\mathfrak{n},G}(\langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle) \oplus O^{\mathfrak{n},G,\ltimes}(\tau^{G'}).$$

Relevant histories. Since our nature policies are restricted to finite probability distributions, we can generate a finite subset of all joint histories that possibly have a non-zero probability at time t given a nature policy θ up to time t. For simplicity, we use the RPOMDP histories. In Appendix F we show that reasoning via POSG histories is equivalent.

Definition 13 (Relevant joint history)**.** *Given a deterministic policy* $\theta^{det} \in \Theta^{det}$*,* $\mathsf{rel} \colon \Theta^{det} \to \mathcal{P}(H^M)$ *gives the set of joint histories which the deterministic policy can reach.*

$$\mathsf{rel}(\theta^{det}) = \{O^M(\langle s_I\rangle)\} \cup \{h \oplus \langle a, u, z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}\rangle \in H^M \mid \theta^{det}(h^{\mathfrak{n}}, a) = u \wedge h \in \mathsf{rel}(\theta^{det})\}.$$

Where $h^{\mathfrak{n}}$ *is the nature observable part of the joint history* h*.*
Given a mixed policy $\theta^{mix} \in \Theta^{mix}$*,* $\mathsf{rel} \colon \Theta^{mix} \to \mathcal{P}(H^M)$ *gives the set of joint histories which the mixed policy can reach. This comes down to the histories that are relevant to one of the deterministic policies the mixed policy randomizes over.*

$$\mathsf{rel}(\theta^{mix}) = \{h \in H^M \mid \exists \theta^{det} \in \Theta^{det}.\, \theta^{mix}(\theta^{det}) > 0 \wedge h \in \mathsf{rel}(\theta^{det})\}.$$

Given a stochastic policy $\theta \in \Theta$*,* $\mathsf{rel} \colon \Theta \to \mathcal{P}(H^M)$ *gives the set of joint histories that the stochastic policy can reach.*

$$\mathsf{rel}(\theta) = \{O^M(\langle s_I\rangle)\} \cup \{h \oplus \langle a, u, z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}\rangle \in H^M \mid \theta(h^{\mathfrak{n}}, a)(u) > 0 \wedge h^{\mathfrak{n}} \in \mathsf{rel}(\theta)\}.$$

Where $h^{\mathfrak{n}}$ *is the nature observable part of the joint history* h*.*

This construction generalizes to relevant nature histories, indicated by $\mathsf{rel}^{\mathfrak{n}}$. We use the sets of relevant histories in our proof of the existence of a Nash equilibrium in Appendix G.

Policy types As introduced in Section 2, stochastic policies map histories to (finite) distributions over actions. Mixed policies, on the other hand, are (finite) distributions over deterministic policies, which in turn map histories to actions deterministically. For convenience, we repeat all types of policies we consider below.

For RPOMDPs:

Stochastic:	$\pi \colon H^{\mathfrak{a},M} \to \Delta(A)$,	$\theta \colon H^{\mathfrak{n},M} \times A \to \Delta(\boldsymbol{U})$,
Deterministic:	$\pi^{det} \colon H^{\mathfrak{a},M} \to A$,	$\theta^{det} \colon H^{\mathfrak{n},M} \times A \to \boldsymbol{U}$,
Mixed:	$\pi^{mix} \in \Delta(H^{\mathfrak{a},M} \to A)$,	$\theta^{mix} \in \Delta(H^{\mathfrak{n},M} \times A \to \boldsymbol{U})$.

And for POSGs:

$$\begin{aligned}
\text{Stochastic:} \quad & \pi\colon H^{\mathfrak{a},G} \to \Delta(\mathcal{A}^{\mathfrak{a}}), && \theta\colon H^{\mathfrak{n},G} \times \mathcal{Z}^{\mathfrak{n}} \to \Delta(\mathcal{A}^{\mathfrak{n}}),\\
\text{Deterministic:} \quad & \pi^{det}\colon H^{\mathfrak{a},G} \to \mathcal{A}^{\mathfrak{a}}, && \theta^{det}\colon H^{\mathfrak{n},G} \times \mathcal{Z}^{\mathfrak{n}} \to \mathcal{A}^{\mathfrak{n}},\\
\text{Mixed:} \quad & \pi^{mix} \in \Delta(H^{\mathfrak{a},G} \to \mathcal{A}^{\mathfrak{a}}), && \theta^{mix} \in \Delta(H^{\mathfrak{n},G} \times \mathcal{Z}^{\mathfrak{n}} \to \mathcal{A}^{\mathfrak{n}}).
\end{aligned}$$

We write $\Pi^{M}, \Pi^{G}, \Pi^{det,M}, \Pi^{det,G}, \Pi^{mix,M}$, and $\Pi^{mix,G}$ for the stochastic, deterministic, and mixed agent policies in RPOMDP and POSG models, respectively. Similarly, we write $\Theta^{M}, \Theta^{G}, \Theta^{det,M}, \Theta^{det,G}, \Theta^{mix,M}$, and $\Theta^{mix,G}$ for the stochastic, deterministic, and mixed nature policies in RPOMDP and POSG models, respectively. Note that we often omit the model indication superscript when it is clear from context in which type of model we operate or the results are equivalent.

The set of deterministic policies can be viewed as a subset of both the set of stochastic and the set of mixed policies using only Dirac distributions. The value function of a mixed policy is computed as follows:

$$\begin{aligned}
V_{\phi}^{\pi^{mix},\theta^{mix}} &= \sum_{\pi^{det} \in \Pi^{det}} \left\{ \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot V_{\phi}^{\pi^{det},\theta^{det}} \right\}\\
&= \sum_{\pi^{det} \in \Pi^{det}} \sum_{\theta^{det} \in \Theta^{det}} \left\{ \pi^{mix}(\pi^{det}) \cdot \theta^{mix}(\theta^{det}) \cdot V_{\phi}^{\pi^{det},\theta^{det}} \right\}.
\end{aligned}$$

A.3 Nature first

Below, we define the same paths to histories mapping for the POSGs of nature first RPOMDPs. The changes to the paths and histories are discussed in Appendix E.

Definition 14 (Paths to joint histories in POSGs)**.**
Similarly, let $O^{G,\ltimes} : Paths^{G,\ltimes} \to H^{G,\ltimes}$ defined by:

$$\begin{aligned}
O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle, a'\rangle) &= \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a'\rangle, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle.\\
O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) &= \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a'\rangle, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle, u, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle, a\rangle.\\
O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle \oplus \tau^{G'}) &= O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) \oplus O^{G,\ltimes}(\tau^{G'}).
\end{aligned}$$

Let $O^{G} : Paths^{G} \to H^{G}$ defined by:

$$\begin{aligned}
O^{G}(\langle\langle s, u^{\hookrightarrow}\rangle, \bot\rangle) &= \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle.\\
O^{G}(\langle\langle s, u^{\hookrightarrow}, \bot\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) &= \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle, u, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle, a\rangle.\\
O^{G}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle \oplus \tau^{G'}) &= O^{G}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) \oplus O^{G,\ltimes}(\tau^{G'}).
\end{aligned}$$

Definition 15 (Paths to agent histories in POSGs)**.**
Similarly, let $O^{\mathfrak{a},G,\ltimes} : Paths^{G,\ltimes} \to H^{\mathfrak{a},G,\ltimes}$ defined by:

$$\begin{aligned}
O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle, a'\rangle) &= \langle\langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle.\\
O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) &= \langle\langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle, a\rangle.\\
O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle \oplus \tau^{G'}) &= O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) \oplus O^{G,\ltimes}(\tau^{G'}).
\end{aligned}$$

Let $O^{\mathfrak{a},G} : Paths^{G} \to H^{\mathfrak{a},G}$ defined by:

$$\begin{aligned}
O^{G}(\langle\langle s, u^{\hookrightarrow}\rangle, \bot\rangle) &= \langle\langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle\rangle.\\
O^{G}(\langle\langle s, u^{\hookrightarrow}, \bot\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) &= \langle\langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle, \langle O^{\mathfrak{a}}_{\bullet}(s), O_{\circ}(s)\rangle, a\rangle.\\
O^{G}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle \oplus \tau^{G'}) &= O^{G}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) \oplus O^{G,\ltimes}(\tau^{G'}).
\end{aligned}$$

Definition 16 (Paths to nature histories in POSGs)**.**
Similarly, let $O^{\mathfrak{n},G,\ltimes} : Paths^{G,\ltimes} \to H^{\mathfrak{n},G,\ltimes}$ defined by:

$$\begin{aligned}
O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}\rangle, a'\rangle) &= \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a'\rangle\rangle.\\
O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) &= \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), a'\rangle, u, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle\rangle.\\
O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle \oplus \tau^{G'}) &= O^{G,\ltimes}(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) \oplus O^{G,\ltimes}(\tau^{G'}).
\end{aligned}$$

Let $O^{\mathfrak{n},G} : Paths^G \to H^{\mathfrak{n},G}$ defined by:

$$O^G(\langle\langle s, u^{\hookrightarrow}\rangle, \bot\rangle) = \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle\rangle.$$
$$O^G(\langle\langle s, u^{\hookrightarrow}, \bot\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) = \langle\langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle, u, \langle O^{\mathfrak{n}}_{\bullet}(s), O_{\circ}(s), \bot\rangle\rangle.$$
$$O^G(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle \oplus \tau^{G'}) = O^G(\langle\langle s, u^{\hookrightarrow}, a'\rangle, u, \langle s, u^{\hookrightarrow}, u\rangle, a\rangle) \oplus O^{G,\ltimes}(\tau^{G'}).$$

Since nature policies no longer depend on the last played agent action in a nature first model, the set of relevant joint histories changes as follows:

Definition 17 (Relevant joint history)**.** *Given a deterministic policy $\theta^{det} \in \Theta^{det}$, $\mathsf{rel}\colon \Theta^{det} \to \mathcal{P}(H^M)$ gives the set of joint histories which the deterministic policy can reach.*

$$\mathsf{rel}(\theta^{det}) = \{O^M(\langle s_I\rangle)\} \cup \{h \oplus \langle a, u, z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}\rangle \in H^M \mid \theta^{det}(h^{\mathfrak{n}}) = u \wedge h \in \mathsf{rel}(\theta^{det})\}.$$

Where $h^{\mathfrak{n}}$ is the nature observable part of the joint history h.
Given a mixed policy $\theta^{mix} \in \Theta^{mix}$, $\mathsf{rel}\colon \Theta^{mix} \to \mathcal{P}(H^M)$ gives the set of joint histories which the mixed policy can reach. This comes down to the histories that are relevant to one of the deterministic policies the mixed policy randomizes over.

$$\mathsf{rel}(\theta^{mix}) = \{h \in H^M \mid \exists \theta^{det} \in \Theta^{det}.\, \theta^{mix}(\theta^{det}) > 0 \wedge h \in \mathsf{rel}(\theta^{det})\}.$$

Given a stochastic policy $\theta \in \Theta$, $\mathsf{rel}\colon \Theta \to \mathcal{P}(H^M)$ gives the set of joint histories that the stochastic policy can reach.

$$\mathsf{rel}(\theta) = \{O^M(\langle s_I\rangle)\} \cup \{h \oplus \langle a, u, z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}\rangle \in H^M \mid \theta(h^{\mathfrak{n}})(u) > 0 \wedge h^{\mathfrak{n}} \in \mathsf{rel}(\theta)\}.$$

Where $h^{\mathfrak{n}}$ is the nature observable part of the joint history h.

This construction, again, generalizes to relevant nature histories, indicated by $\mathsf{rel}^{\mathfrak{n}}$. The sets of relevant histories for nature first RPOMDPs are needed to adjust the proof of the existence of a Nash equilibrium in Appendix G to the nature first setting.

B From General RPOMDP to RPOMDP With Deterministic Observations

This appendix shows that our definition of RPOMDPs with deterministic state-based observations is non-restrictive.

Similar to in [Chatterjee *et al.*, 2016], we can transform an RPOMDP with a stochastic or uncertain observation function into an equivalent one with a deterministic observation function. Let $M = (S, A, \boldsymbol{TO}, R, Z_\bullet^{\mathfrak{a}}, Z_\bullet^{\mathfrak{n}}, Z_\circ)$ be an RPOMDP with an *uncertain transition observation function* defined as $\boldsymbol{TO} \colon \boldsymbol{U} \to (S \times A \to \Delta(S \times Z_\bullet^{\mathfrak{a}} \times Z_\bullet^{\mathfrak{n}} \times Z_\circ))$. Note that this definition combines the transition and observation functions into one to allow for more intricate dependencies. Multiple independent functions can replace the $\boldsymbol{TO}$ function. This does not change the transformation below.

From the RPOMDP M, we construct a larger, equivalent, RPOMDP $M' = (S', A, \boldsymbol{T}, R', Z_\bullet^{\mathfrak{a}}, Z_\bullet^{\mathfrak{n}}, Z_\circ, O_\bullet^{\mathfrak{a}}, O_\bullet^{\mathfrak{n}}, O_\circ)$ with $S' = S \times Z_\bullet^{\mathfrak{a}} \times Z_\bullet^{\mathfrak{n}} \times Z_\circ$, adjusting the reward function according to the new state space $R' \colon S' \times A \to \mathbb{R}$. We split the original transition observation function $\boldsymbol{TO}$ in a transition function $\boldsymbol{T} \colon \boldsymbol{U} \to (S' \times A \to \Delta(S'))$ and three separate deterministic observation functions $O_\bullet^{\mathfrak{a}\prime} \colon S' \to Z_\bullet^{\mathfrak{a}}$,$O_\bullet^{\mathfrak{n}\prime} \colon S' \to Z_\bullet^{\mathfrak{n}}$, and ${O_\circ}' \colon S' \to Z_\circ$. The functions are then defined as follows:

- $\boldsymbol{T}(u)(\langle s, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \rangle, a, \langle s', z_\bullet^{\mathfrak{a}\prime}, z_\bullet^{\mathfrak{n}\prime}, {z_\circ}' \rangle) = \boldsymbol{TO}(u)(s, a, s', z_\bullet^{\mathfrak{a}\prime}, z_\bullet^{\mathfrak{n}\prime}, {z_\circ}')$,
- $R'(\langle s, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \rangle, a) = R(s, a)$,
- $O_\bullet^{\mathfrak{a}}(\langle s, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \rangle) = z_\bullet^{\mathfrak{a}}$,
- $O_\bullet^{\mathfrak{n}}(\langle s, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \rangle) = z_\bullet^{\mathfrak{n}}$,
- $O_\circ(\langle s, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \rangle) = z_\circ$.

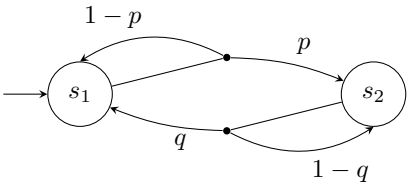
The arbitrary RPOMDP M has now been transformed into an RPOMDP M' that satisfies our Definition 3, showing that our assumption of deterministic state-based observations is indeed non-restrictive.

C Stickiness Examples

This appendix contains three examples of a stickiness function: zero, full, and observation-based stickiness.

C.1 Zero and Full Stickiness

As mentioned in Section 3.1, zero and full stickiness are the extremes of the spectrum of stickiness types where nature's choices never or always stick, respectively. We revisit the RMDP in Figure 1 in the main text and discuss the zero and full stickiness interpretations of the model.



$\boldsymbol{U}^1 = \{p \in [0.1, 0.9], q \in [0.1, 0.9]\}$
$\boldsymbol{U}^2 = \{p \in [0.1, 0.4], q = 2p\}$

Figure 1: An example RMDP with two uncertainty sets.

Example 3 (Stickiness). Consider the RMDP in Figure 1 and uncertainty set $\boldsymbol{U}^1$. We interpret this RMDP as an RPOMDP with full observability for both players. Regardless of the stickiness, at the start of the game, nature has to choose a variable assignment $u \in \boldsymbol{U}^1$. Under full stickiness, the rest of the game is now determined, as nature can only choose the same values for p and q as in the initial variable assignment. If we assume zero stickiness, then at the next state and all future states, whether s_1 or s_2, nature can choose any new variable assignment $u' \in \boldsymbol{U}^1$.

C.2 Observation-Based Stickiness

We allow a stickiness function to depend on what nature observes, *i.e.*, its private observations $Z_\bullet^{\mathrm{n}}$, public observations $Z_\circ$, and the agent's actions A. To define such stickiness functions, we denote for each variable $v \in U$ the state-action pairs it influences by $v^{s,a} \colon S \times A \to \{0, 1\}$.

An example of an observation-based stickiness function is:

$$\begin{aligned}&\forall v \in U, z_\bullet^{\mathrm{n}} \in Z_\bullet^{\mathrm{n}}, z_\circ \in Z_\circ, a \in A.\\ &\quad \mathsf{stick}(v, z_\bullet^{\mathrm{n}}, z_\circ, a) = 1 \iff \exists s \in S. O_\bullet^{\mathrm{n}}(s) = z_\bullet^{\mathrm{n}} \wedge O_\circ(s) = z_\circ \wedge v^{s,a}(s, a) = 1.\end{aligned}$$

Under observation-based stickiness, a variable only sticks if there is a possibility that it influenced the actual transition based on the last observations and actions. This means that all variables that influence both a state with observations $z_\bullet^{\mathrm{n}}, z_\circ$ and the chosen action a stick. The intuition behind observation-based stickiness is that nature only needs to optimize for the transitions it might influence at that given point in time. Note that this stickiness does not take the entire history into account, so there can still be restrictions on variables that nature knows cannot influence the actual transition, as can be seen in the discussion of the right example below.

$\boldsymbol{U} = \{p \in [0.1, 0.9], q \in [0.1, 0.9]\}$

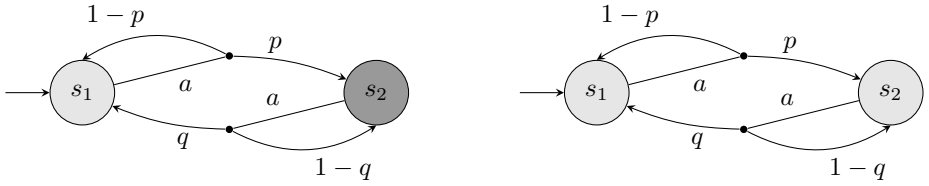


Figure 6: Two example RPOMDPs with the same uncertainty set.

Example 4. Figure 6 depicts two RPOMDPs. For simplicity, these RPOMDPs have no private observations. Furthermore, we interpret these RPOMDPs with the agent first semantics. Note that the left RPOMDP corresponds to the fully observable RPOMDP interpretation of the RMDP in Figure 1. For both RPOMDPs, we have that:

$$p^{s,a}(s_1, a) = 1, p^{s,a}(s_2, a) = 0, q^{s,a}(s_1, a) = 0, q^{s,a}(s_2, a) = 1.$$

So variable p only influences transitions from state s_1, and variable q only influences transitions from state s_2. Like in Example 3, in both RPOMDPs, nature has to choose a variable assignment $u \in \boldsymbol{U}$ at the start of the game. First, looking at

the left RPOMDP and assuming observation-based stickiness, the variable p becomes restricted after this initial choice, since $O_\circ(s_1) = \bigcirc \wedge p^{s,a}(s_1, a) = 1$, so $\mathsf{stick}(p, \bot, \bigcirc, a) = 1$. Variable q remains unrestricted, as state s_2, the only state that q influences, has a different observation. As long as the agent remains in state s_1, nature may choose different assignments for q. Once the agent reaches state s_2, whichever value it assigns to q next will stick, and from then on, the game is fully determined as also with full stickiness.

Looking at the right RPOMDP, again assuming observation-based stickiness, both variables immediately become restricted after the initial choice. p again becomes restricted because $O_\circ(s_1) = \bigcirc \wedge p^{s,a}(s_1, a) = 1$, so we have $\mathsf{stick}(p, \bot, \bigcirc, a) = 1$. Now, since $s2$ has the same observation as s_1, we have $O_\circ(s2) = \bigcirc \wedge q^{s,a}(s_2, a) = 1$, which gives us $\mathsf{stick}(q, \bot, \bigcirc, a) = 1$.

D Uncertainty Assumptions Matter

In this appendix, we elaborate on the results established in Theorem 1 and its proof.

Theorem 1 (Uncertainty assumptions matter)**.** *For an RPOMDP M, let $V_{fh}^{*,M}$ 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 $V_{fh}^{*,M_1} \neq V_{fh}^{*,M_2}$,*
2. *There exist RPOMDPs M_1, M_2 that only differ in their order of play, such that $V_{fh}^{*,M_1} \neq V_{fh}^{*,M_2}$.*

In the following four subsections, we compare tuples of R(PO)MDPs which only differ in either the stickiness or the order of play. The first two subsections focus on differences in the stickiness, and the latter two focus on differences in the order of play:

1. Full stickiness versus zero stickiness in an (s, a)-rectangular model (Appendix D.1).
2. Full stickiness versus observation-based stickiness versus zero stickiness in an a-rectangular model (Appendix D.2).
3. Agent first versus nature first in a simple model (Appendix D.3).
4. Agent first versus nature first in an a-rectangular full sticky model (Appendix D.4).

For each of the tuples of RPOMDPs, we first state the value functions given agent and nature policies. In principle, we use the sets of mixed nature policies Θ^{mix} for computing the optimal value and policy, which are equivalent to the original sets of stochastic policies, as shown in Appendix G.3. However, in three of the four RPOMDP tuples, we can consider deterministic nature policies instead of mixed nature policies, as discussed in more detail in Appendix D.1. For legibility reasons, we switch back to the stochastic policy notation when writing the optimal policy.

Given the value functions and the types of optimal policies, we used the 3D calculator of [GeoGebra GmbH, 2024] to search for the optimal value. We plotted the value functions for finding the optimal agent and nature policies separately. In both cases, we look for the policy values that optimize the worst-case scenario from the player's perspective. Once we found the value that both players can achieve regardless of the other player's policy, we found the Nash equilibrium value and policies. The models used for finding the optimal values and policies can be found at https://github.com/LAVA-LAB/RPOMDP_game_semantics_value_functions.

The computed optimal values show that the optimal values differ between the RPOMDPs in the tuples. These tuples of RPOMDPs therefore prove Theorem 1.

After the optimal value computation, we discuss the structural differences of the equivalent POSGs. Although the structural differences between two POSGs do not ensure that they have a different optimal value, the differences do provide an intuition in the cases where we know that the optimal values differ.

D.1 Stickiness Matters

We first revisit the RPOMDP in Figure 2 and show how we computed the optimal values to show that stickiness matters in RPOMDPs.

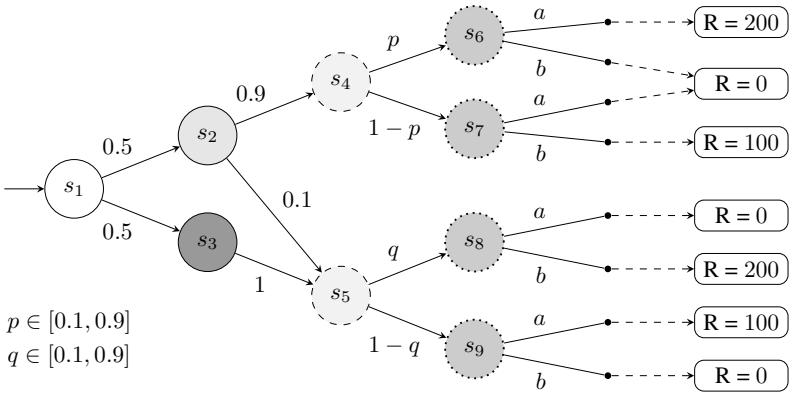


Figure 2: An RPOMDP where full and zero stickiness do not coincide in their optimal value.

For notation purposes, we use the distinguishing second observation to identify the longer history inputs for policies where needed. For $\pi \in \Pi$, we write $\pi^{\circ} = \pi(\bigcirc\bigcirc\cdot\cdot)$ and $\pi^{\bullet} = \pi(\bigcirc\bullet\cdot\cdot)$. For $\theta \in \Theta$ of the full stickiness RPOMDP, we write

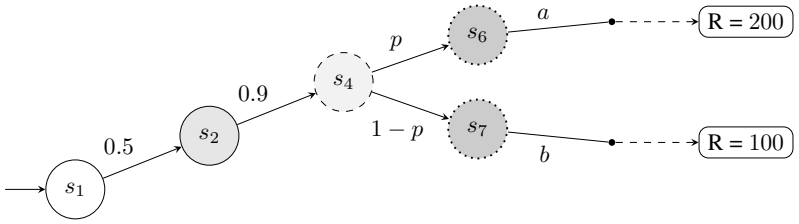
$\theta^{\circ} = \theta(\circ)$, and for $\theta \in \Theta$ of the zero stickiness RPOMDP, we write $\theta^{\circ} = \theta(\circ\circ)$ and $\theta^{\bullet} = \theta(\circ\bullet)$. Using this notation, we can write the value function for the full stickiness RPOMDP M_1 in Figure 2 as:

$$V_{\text{fh}}^{M_1}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(0.5 \cdot 0.9 \cdot \big(\theta^{det,\circ}(p) \cdot \pi^{\circ}(a) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot \pi^{\circ}(b) \cdot 100\big) \\ + 0.5 \cdot 0.1 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\circ}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \cdot \pi^{\circ}(a) \cdot 100\big) \\ + 0.5 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\bullet}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \cdot \pi^{\bullet}(a) \cdot 100\big) \Big)$$

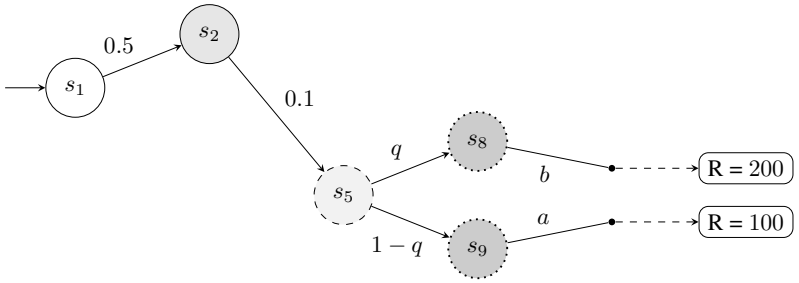
And for the zero stickiness RPOMDP M_2:

$$V_{\text{fh}}^{M_2}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(0.5 \cdot 0.9 \cdot \big(\theta^{det,\circ}(p) \cdot \pi^{\circ}(a) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot \pi^{\circ}(b) \cdot 100\big) \\ + 0.5 \cdot 0.1 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\circ}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \cdot \pi^{\circ}(a) \cdot 100\big) \\ + 0.5 \cdot \big(\theta^{det,\bullet}(q) \cdot \pi^{\bullet}(b) \cdot 200 + (1 - \theta^{det,\bullet}(q)) \cdot \pi^{\bullet}(a) \cdot 100\big) \Big).$$

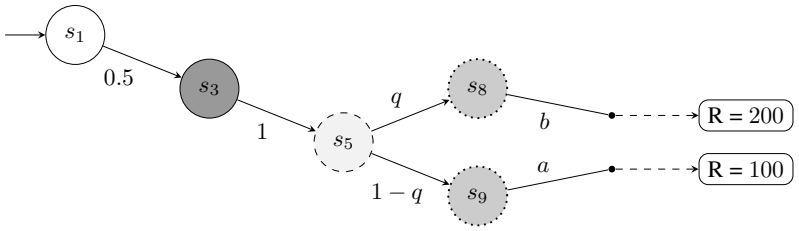
We construct these value functions by following all possible paths leading to non-zero rewards. We combined some of the paths to keep the formula manageable. The three bigger terms correspond to the three subparts of the RPOMDP in Figure 2 shown in Figure 7. A multiplication corresponds to successive branches, whereas addition corresponds to parallel branches. Note that we removed the paths leading to rewards of zero. Also note that in this case, we used the full stickiness theta. For the zero stickiness theta, the history on which choices are based is different, but the multiplications, additions, and how we wrote down the formula still correspond to the subparts below.



(a) Subpart of the RPOMDP responsible for $0.5 \cdot 0.9 \cdot \big(\theta^{det,\circ}(p) \cdot \pi^{\circ}(a) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot \pi^{\circ}(b) \cdot 100\big)$



(b) Subpart of the RPOMDP responsible for $0.5 \cdot 0.1 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\circ}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \cdot \pi^{\circ}(a) \cdot 100\big)$



(c) Subpart of the RPOMDP responsible for $0.5 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\bullet}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \cdot \pi^{\bullet}(a) \cdot 100\big)$

Figure 7: Representation of the split used in the construction of value function of the RPOMDP in Figure 2.

We show that we can reason with deterministic nature policies by showing that the value functions are linear in the deterministic nature policies. Combining this with the convexity of the uncertainty sets, we show that any finite probability distribution over

the deterministic nature policies, *i.e.*, any mixed nature policy, can be rewritten as another deterministic nature policy that is contained in the policy set. We can, therefore, limit ourselves to searching for the optimal policy in the deterministic nature policy set. We first prove the following lemma:

Lemma 1. *Given the full stickiness and zero stickiness RPOMDPs M_1 and M_2 of Figure 2:*

$$\forall \theta^{mix} \in \Theta^{mix}, \exists \theta^{det} \in \Theta^{det}, \forall \pi \in \Pi.\ V_{fh}^{M_1}(\pi, \theta^{mix}) = V_{fh}^{M_1}(\pi, \theta^{det}),$$

$$\forall \theta^{mix} \in \Theta^{mix}, \exists \theta^{det} \in \Theta^{det}, \forall \pi \in \Pi.\ V_{fh}^{M_2}(\pi, \theta^{mix}) = V_{fh}^{M_2}(\pi, \theta^{det}).$$

Proof. We focus on the full stickiness RPOMDP M_1. The result for the zero stickiness RPOMDP M_2 follows from the same steps. Let $\theta^{mix} \in \Theta^{mix}$ be an arbitrary mixed nature policy and $\pi \in \Pi$ an arbitrary stochastic agent policy. Then, we can compute the value as follows:

$$\begin{aligned} V_{\text{fh}}^{M_1}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 0.5 \cdot 0.9 \cdot \big(\theta^{det,\circ}(p) \cdot \pi^{\circ}(a) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot \pi^{\circ}(b) \cdot 100\big) \\ & + 0.5 \cdot 0.1 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\circ}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \cdot \pi^{\circ}(a) \cdot 100\big) \\ & + 0.5 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\bullet}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \cdot \pi^{\bullet}(a) \cdot 100\big)\Big). \end{aligned}$$

Simplify:

$$\begin{aligned} = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 90 \cdot \pi^{\circ}(a) \cdot \theta^{det,\circ}(p) + 45 \cdot \pi^{\circ}(b) - 45 \cdot \pi^{\circ}(b) \cdot \theta^{det,\circ}(p) \\ & + 10 \cdot \pi^{\circ}(b) \cdot \theta^{det,\circ}(q) + 5 \cdot \pi^{\circ}(a) - 5 \cdot \pi^{\circ}(a) \cdot \theta^{det,\circ}(q) \\ & + 100 \cdot \pi^{\bullet}(b) \cdot \theta^{det,\circ}(q) + 50 \cdot \pi^{\bullet}(a) - 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\circ}(q)\Big). \end{aligned}$$

By definition of probability distributions: $\sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) = 1$, so we can move the terms depending only on π out of the summation:

$$\begin{aligned} & = 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) + \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big((90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p) \\ & \quad + (10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a) + 100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\circ}(q)\Big). \end{aligned}$$

Split the summation:

$$\begin{aligned} & = 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) + \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big((90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p)\Big) \\ & \quad + \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big((10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a) + 100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\circ}(q)\Big). \end{aligned}$$

Move the multiplication terms only depending on π out of the summations:

$$\begin{aligned} & = 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\ & \quad + (90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \theta^{det,\circ}(p) \\ & \quad + (10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a) + 100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \theta^{det,\circ}(q). \end{aligned}$$

The uncertainty set $\boldsymbol{U}$ of M_1 is convex. We, therefore, know:

$$\forall u, u' \in \boldsymbol{U}, \forall \alpha \in [0,1].\ \alpha u + (1-\alpha) u' \in \boldsymbol{U}.$$

By definition of valid policies (Section 3.1), we know:

$$\forall \theta^{det} \in \Theta^{det}, \forall h \in H^{\mathfrak{n}}, \forall a \in A.\ \theta^{det}(h, a) \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(h)).$$

At the non-singleton point of choice for the full stickiness nature policies, *i.e.*, history $\circ$, $\mathsf{fix}(\circ) = \emptyset = u^{\perp}$. We, therefore, know:

$$\forall \theta^{det} \in \Theta^{det}.\ \theta^{det,\circ} \in \boldsymbol{U}^{\mathcal{P}}(u^{\perp}) = \boldsymbol{U}.$$

So we can create nature policy θ'^{det} for which:

$$\theta'^{det,\circ}(p) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \theta^{det,\circ}(p),$$
$$\theta'^{det,\circ}(q) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \theta^{det,\circ}(q).$$

Since at the only non-singleton point of choice, θ'^{det} is a convex combination of elements of a convex set, we know that:

$$\theta'^{det,\circ} \in \boldsymbol{U} = \boldsymbol{U}^{\mathcal{P}}(u^{\perp}).$$

As all other choices are singletons, we have that:

$$\forall h \in H^{\mathfrak{n}}, \forall a \in A.\ \theta'^{det}(h,a) \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(h)).$$

From which we can conclude that θ'^{det} is a valid deterministic nature policy, *i.e.*, $\theta'^{det} \in \Theta^{det}$. We continue by showing that $V_{\text{fh}}^{M_1}(\pi, \theta'^{det}) = V_{\text{fh}}^{M_1}(\pi, \theta^{mix})$ for an arbitrary agent policy $\pi \in \Pi$.

$$\begin{aligned} V_{\text{fh}}^{M_1}(\pi, \theta'^{det}) &= 0.5 \cdot 0.9 \cdot \big(\theta'^{det,\circ}(p) \cdot \pi^{\circ}(a) \cdot 200 + (1 - \theta'^{det,\circ}(p)) \cdot \pi^{\circ}(b) \cdot 100\big) \\ &\quad + 0.5 \cdot 0.1 \cdot \big(\theta'^{det,\circ}(q) \cdot \pi^{\circ}(b) \cdot 200 + (1 - \theta'^{det,\circ}(q)) \cdot \pi^{\circ}(a) \cdot 100\big) \\ &\quad + 0.5 \cdot \big(\theta'^{det,\circ}(q) \cdot \pi^{\bullet}(b) \cdot 200 + (1 - \theta'^{det,\circ}(q)) \cdot \pi^{\bullet}(a) \cdot 100\big). \end{aligned}$$

Simplify:

$$\begin{aligned} &= 90 \cdot \pi^{\circ}(a) \cdot \theta'^{det,\circ}(p) + 45 \cdot \pi^{\circ}(b) - 45 \cdot \pi^{\circ}(b) \cdot \theta'^{det,\circ}(p) \\ &\quad + 10 \cdot \pi^{\circ}(b) \cdot \theta'^{det,\circ}(q) + 5 \cdot \pi^{\circ}(a) - 5 \cdot \pi^{\circ}(a) \cdot \theta'^{det,\circ}(q) \\ &\quad + 100 \cdot \pi^{\bullet}(b) \cdot \theta'^{det,\circ}(q) + 50 \cdot \pi^{\bullet}(a) - 50 \cdot \pi^{\bullet}(a) \cdot \theta'^{det,\circ}(q). \end{aligned}$$

Reorder:

$$\begin{aligned} &= 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\ &\quad + (90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta'^{det,\circ}(p) \\ &\quad + (10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a) + 100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta'^{det,\circ}(q). \end{aligned}$$

Using the definition of $\theta'^{det,\circ}$:

$$\begin{aligned} &= 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\ &\quad + (90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \theta^{det,\circ}(p) \\ &\quad + (10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a) + 100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \theta^{det,\circ}(q) \\ &= V_{\text{fh}}^{M_1}(\pi, \theta^{mix}). \end{aligned}$$

Now we have that:

$$\forall \pi \in \Pi.\ V_{\text{fh}}^{M_1}(\pi, \theta^{mix}) = V_{\text{fh}}^{M_1}(\pi, \theta'^{det}),$$

So we can conclude that:

$$\forall \theta^{mix} \in \Theta^{mix}, \exists \theta^{det} \in \Theta^{det}, \forall \pi \in \Pi.\ V_{\text{fh}}^{M_1}(\pi, \theta^{mix}) = V_{\text{fh}}^{M_1}(\pi, \theta^{det}).$$

□

We can now prove the following theorem:

Proposition 3. *Given the full stickiness and zero stickiness RPOMDPs M_1 and M_2 of Figure 2:*

$$\sup_{\pi \in \Pi} \inf_{\theta^{mix} \in \Theta^{mix}} V_{\mathit{fh}}^{M_1}(\pi, \theta^{mix}) = \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} V_{\mathit{fh}}^{M_1}(\pi, \theta^{det}),$$

$$\sup_{\pi \in \Pi} \inf_{\theta^{mix} \in \Theta^{mix}} V_{\mathit{fh}}^{M_2}(\pi, \theta^{mix}) = \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} V_{\mathit{fh}}^{M_2}(\pi, \theta^{det}).$$

Proof. For both M_1 and M_2, the $\leq$ direction directly follows from Lemma 1 and the $\geq$ direction from $\Theta^{det} \subseteq \Theta^{mix}$. □

Using Proposition 3, we can focus on deterministic nature policies in our computation of the optimal value function. We can compute the optimal value for the full stickiness model as follows:

$$\begin{aligned} V_{\text{fh}}^{*,M_1} = \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} \big\{ & 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\ & + (90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p) \\ & + (10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a) + 100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\circ}(q) \big\}. \end{aligned}$$

And for the zero stickiness model, simplified in the same manner as the full stickiness value function:

$$\begin{aligned} V_{\text{fh}}^{*,M_2} = \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} \big\{ & 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\ & + (90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p) \\ & + (10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a)) \cdot \theta^{det,\circ}(q) \\ & + (100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\bullet}(q) \big\}. \end{aligned}$$

Since $\theta^{det,\circ}$ is independent from $\theta^{det,\bullet}$ and π° is independent from $\pi^{\bullet}$, we can rewrite the zero stickiness optimal value function as:

$$\begin{aligned} V_{\text{fh}}^{*,M_2} = & \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} \big\{ 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) \\ & + (90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p) \\ & + (10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a)) \cdot \theta^{det,\circ}(q) \big\} \\ & + \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} \big\{ 50 \cdot \pi^{\bullet}(a) + (100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\bullet}(q) \big\}. \end{aligned}$$

Table 3 displays the computed optimal values and policies, showing the differences between the full and zero stickiness assumptions. An underscore indicates that the value assigned to this variable does not influence the optimal value of the RPOMDP.

	Full stickiness	Zero stickiness
Optimal value	$66\frac{2}{3}$	$65\frac{1}{2}$
Optimal agent policy	○○◔◔ $\mapsto \{a \mapsto \frac{1}{3}, b \mapsto \frac{2}{3}\}$, ○●◔◔ $\mapsto \{a \mapsto \frac{7}{10}, b \mapsto \frac{3}{10}\}$	○○◔◔ $\mapsto \{a \mapsto \frac{1}{3}, b \mapsto \frac{2}{3}\}$, ○●◔◔ $\mapsto \{a \mapsto \frac{2}{3}, b \mapsto \frac{1}{3}\}$
Optimal nature policy	○ $\mapsto \{p \mapsto \frac{1}{3}, q \mapsto \frac{1}{3}\}$	○○◔ $\mapsto \{p \mapsto \frac{83}{270}, q \mapsto \frac{1}{10}\}$, ○●◔ $\mapsto \{p \mapsto \_, q \mapsto \frac{1}{3}\}$

Table 3: Optimal values and policies for the full stickiness and zero stickiness interpretations of the RPOMDP in Figure 2.

Underlying POSGs

Figure 4 (restated below) depicts the first couple of states of the full and zero stickiness POSGs of the RPOMDP in Figure 2. We briefly discuss the structural differences. In the zero stickiness case, we have an infinite action choice for nature at every nature state, but every choice will reach the same unrestricted agent states, as variable assignments never stick in the zero stickiness case. Due to this, the state space of the zero stickiness POSG is finite, consisting of $S + S \times A$ states, where S is the set of states and A is the set of actions of the original RPOMDP. This can be seen in Figure 4 at nature state $\langle s_1, \{\}, \bot \rangle$, where each choice reaches the same agent states $\langle s_2, \{\} \rangle$ and $\langle s_3, \{\} \rangle$. At this point, the choice does not influence the transition probability, so all choices go to the agent states with exactly the same probability and essentially have no influence. This is no longer the case at nature states $\langle s_4, \{\}, \bot \rangle$ and $\langle s_5, \{\}, \bot \rangle$, where the values chosen for p and q directly influence the probability of reaching agent state $\langle s_6, \{\} \rangle$, $\langle s_7, \{\} \rangle$, $\langle s_8, \{\} \rangle$, and $\langle s_9, \{\} \rangle$.

In the full stickiness case, on the other hand, each of the infinite action choices at the first nature state $\langle s_1, \{\}, \bot \rangle$ leads to a *unique* continuation of the POSG, as each game continues to agent states with different restrictions on nature's choice. The full stickiness case, hence, has an infinite state space but only one infinite action choice, namely the first.

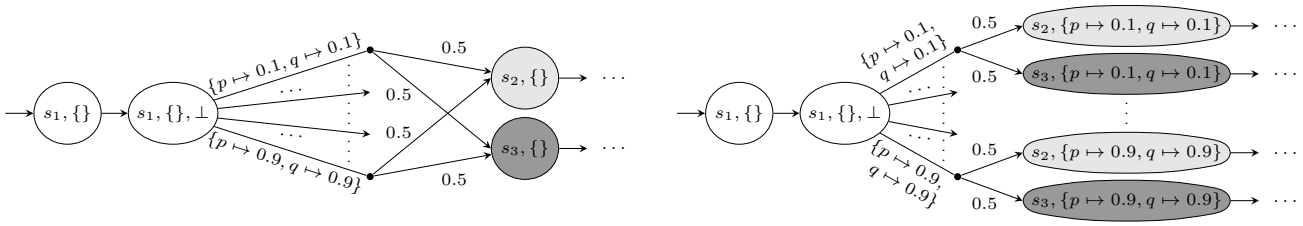


Figure 4: First states of zero stickiness (left) and full stickiness (right) POSGs of the RPOMDP in Figure 2.

D.2 Observation-based stickiness

Next, we look at the RPOMDP in Figure 8 to show that observation-based stickiness also differs in value from full and zero stickiness. Note that this model extends the RPOMDP in Figure 2. We interpret this model with nature first semantics.

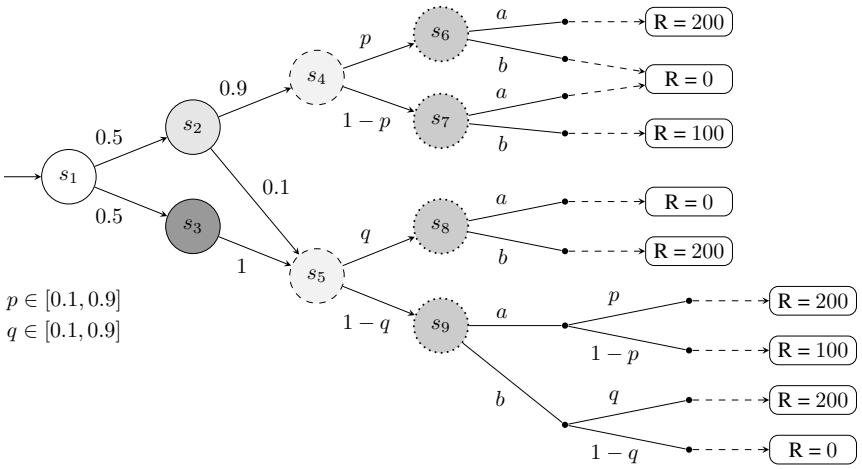


Figure 8: An RPOMDP where observation-based stickiness also leads to different values.

For $\pi \in \Pi$, we write $\pi^{\circ} = \pi(\bigcirc\bigcirc(\cdot)(\cdot))$ and $\pi^{\bullet} = \pi(\bigcirc\bullet(\cdot)(\cdot))$. Similarly, for $\theta \in \Theta$ of the full stickiness RPOMDP, we write $\theta^{\circ} = \theta(\bigcirc)$, for $\theta \in \Theta$ of the observation-based stickiness RPOMDP, we write $\theta^{\circ} = \theta(\bigcirc\bigcirc(\cdot))$ and $\theta^{\bullet} = \theta(\bigcirc\bullet(\cdot))$, and for $\theta \in \Theta$ of the zero stickiness RPOMDP, we write $\theta^{\circ} = \theta(\bigcirc\bigcirc(\cdot))$, $\theta^{\bullet} = \theta(\bigcirc\bullet(\cdot))$, $\theta^{\circ\cdot} = \theta(\bigcirc\bullet(\cdot)(\cdot))$, and $\theta^{\bullet\cdot} = \theta(\bigcirc\bullet(\cdot)(\cdot))$. Using this notation, we can construct the value functions for the various stickiness interpretations of the RPOMDP in Figure 2. We construct these value functions following the same approach as for the value functions of the RPOMDP in Figure 2, see Appendix D.1 and Figure 7. The value function for the full stickiness RPOMDP M_1 is:

$$
\begin{aligned}
V_{\text{fh}}^{M_1}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 0.5 \cdot 0.9 \cdot \big(\theta^{det,\circ}(p) \cdot \pi^{\circ}(a) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot \pi^{\circ}(b) \cdot 100\big) \\
& + 0.5 \cdot 0.1 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\circ}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \\
& \quad \cdot (\pi^{\circ}(a) \cdot (\theta^{det,\circ}(p) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot 100) \\
& \quad + \pi^{\circ}(b) \cdot \theta^{det,\circ}(q) \cdot 200)\big) \\
& + 0.5 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\bullet}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \\
& \quad \cdot (\pi^{\bullet}(a) \cdot (\theta^{det,\circ}(p) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot 100) \\
& \quad + \pi^{\bullet}(b) \cdot \theta^{det,\circ}(q) \cdot 200)\big)\Big).
\end{aligned}
$$

Which simplifies to:

$$
\begin{aligned}
= \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\
& + (95 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b) + 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\circ}(p) \\
& + (20 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a) + 200 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\circ}(q) \\
& - (5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\circ}(p) \cdot \theta^{det,\circ}(q) \\
& - (10 \cdot \pi^{\circ}(b) + 100 \cdot \pi^{\bullet}(b)) \cdot \theta^{det,\circ}(q)^2 \Big).
\end{aligned}
$$

This function is not linear in the deterministic nature policies, as quickly follows from the multiplication of θ terms. We, therefore, need to search for the optimal nature policy in the set of mixed nature policies.

We can similarly write the value function for the observation-based stickiness RPOMDP M_2:

$$
\begin{aligned}
V_{\text{fh}}^{M_2}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 0.5 \cdot 0.9 \cdot \big(\theta^{det,\circ}(p) \cdot \pi^{\circ}(a) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot \pi^{\circ}(b) \cdot 100\big) \\
& + 0.5 \cdot 0.1 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\circ}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \\
& \quad \cdot (\pi^{\circ}(a) \cdot (\theta^{det,\circ}(p) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot 100) \\
& \quad + \pi^{\circ}(b) \cdot \theta^{det,\circ}(q) \cdot 200)\big) \\
& + 0.5 \cdot \big(\theta^{det,\bullet}(q) \cdot \pi^{\bullet}(b) \cdot 200 + (1 - \theta^{det,\bullet}(q)) \\
& \quad \cdot (\pi^{\bullet}(a) \cdot (\theta^{det,\bullet}(p) \cdot 200 + (1 - \theta^{det,\bullet}(p)) \cdot 100) \\
& \quad + \pi^{\bullet}(b) \cdot \theta^{det,\bullet}(q) \cdot 200)\big)\Big).
\end{aligned}
$$

Which simplifies to:

$$
\begin{aligned}
= \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\
& + (95 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p) \\
& + (20 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a)) \cdot \theta^{det,\circ}(q) \\
& - 5 \cdot \pi^{\circ}(a) \cdot \theta^{det,\circ}(p) \cdot \theta^{det,\circ}(q) \\
& - 10 \cdot \pi^{\circ}(b) \cdot \theta^{det,\circ}(q)^2 \\
& + 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet}(p) \\
& + (200 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\bullet}(q) \\
& - 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet}(p) \cdot \theta^{det,\bullet}(q) \\
& - 100 \cdot \pi^{\bullet}(b) \cdot \theta^{det,\bullet}(q)^2 \Big).
\end{aligned}
$$

And the value function for the zero stickiness RPOMDP M_3:

$$
\begin{aligned}
V_{\text{fh}}^{M_3}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 0.5 \cdot 0.9 \cdot \big(\theta^{det,\circ}(p) \cdot \pi^{\circ}(a) \cdot 200 + (1 - \theta^{det,\circ}(p)) \cdot \pi^{\circ}(b) \cdot 100\big) \\
& + 0.5 \cdot 0.1 \cdot \big(\theta^{det,\circ}(q) \cdot \pi^{\circ}(b) \cdot 200 + (1 - \theta^{det,\circ}(q)) \\
& \quad \cdot (\pi^{\circ}(a) \cdot (\theta^{det,\circlearrowleft}(p) \cdot 200 + (1 - \theta^{det,\circlearrowleft}(p)) \cdot 100) \\
& \quad + \pi^{\circ}(b) \cdot \theta^{det,\circlearrowleft}(q) \cdot 200)\big) \\
& + 0.5 \cdot \big(\theta^{det,\bullet}(q) \cdot \pi^{\bullet}(b) \cdot 200 + (1 - \theta^{det,\bullet}(q)) \\
& \quad \cdot (\pi^{\bullet}(a) \cdot (\theta^{det,\circlearrowleft}(p) \cdot 200 + (1 - \theta^{det,\circlearrowleft}(p)) \cdot 100) \\
& \quad + \pi^{\bullet}(b) \cdot \theta^{det,\circlearrowleft}(q) \cdot 200)\big)\Big).
\end{aligned}
$$

Which simplifies to:

$$
\begin{aligned}
= \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\
& + (90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p) \\
& + (10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a)) \cdot \theta^{det,\circ}(q) \\
& + 5 \cdot \pi^{\circ}(a) \cdot \theta^{det,\circ\!\!\cdot\!\!\circ}(p) \\
& + 10 \cdot \pi^{\circ}(b) \cdot \theta^{det,\circ\!\!\cdot\!\!\circ}(q) \\
& - 5 \cdot \pi^{\circ}(a) \cdot \theta^{det,\circ}(q) \cdot \theta^{det,\circ\!\!\cdot\!\!\circ}(p) \\
& - 10 \cdot \pi^{\circ}(b) \cdot \theta^{det,\circ}(q) \cdot \theta^{det,\circ\!\!\cdot\!\!\circ}(q) \\
& + (100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\bullet}(q) \\
& + 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet\!\!\cdot\!\!\bullet}(p) \\
& + 100 \cdot \pi^{\bullet}(b) \cdot \theta^{det,\bullet\!\!\cdot\!\!\bullet}(q) \\
& - 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet}(q) \cdot \theta^{det,\bullet\!\!\cdot\!\!\bullet}(p) \\
& - 100 \cdot \pi^{\bullet}(b) \cdot \theta^{det,\bullet}(q) \cdot \theta^{det,\bullet\!\!\cdot\!\!\bullet}(q) \Big).
\end{aligned}
$$

Using the above value functions, we can compute the optimal value for the full stickiness model as follows:

$$
\begin{aligned}
V_{\text{fh}}^{*,M_1} = \sup_{\pi \in \Pi} \inf_{\theta^{mix} \in \Theta^{mix}} \Big\{ \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\
& + (95 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b) + 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\circ}(p) \\
& + (20 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a) + 200 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\circ}(q) \\
& - (5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\circ}(p) \cdot \theta^{det,\circ}(q) \\
& - (10 \cdot \pi^{\circ}(b) + 100 \cdot \pi^{\bullet}(b)) \cdot \theta^{det,\circ}(q)^2 \Big) \Big\}.
\end{aligned}
$$

And the optimal value for the observation-based stickiness model:

$$
\begin{aligned}
V_{\text{fh}}^{*,M_2} = \sup_{\pi \in \Pi} \inf_{\theta^{mix} \in \Theta^{mix}} \Big\{ \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\
& + (95 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p) \\
& + (20 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a)) \cdot \theta^{det,\circ}(q) \\
& - 5 \cdot \pi^{\circ}(a) \cdot \theta^{det,\circ}(p) \cdot \theta^{det,\circ}(q) \\
& - 10 \cdot \pi^{\circ}(b) \cdot \theta^{det,\circ}(q)^2 \\
& + 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet}(p) \\
& + (200 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\bullet}(q) \\
& - 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet}(p) \cdot \theta^{det,\bullet}(q) \\
& - 100 \cdot \pi^{\bullet}(b) \cdot \theta^{det,\bullet}(q)^2 \Big) \Big\}.
\end{aligned}
$$

As histories ○○(○)⁂ and ○●(○)⁂, and ○○(○) and ○●(○) are mutually exclusive, and the related non-singleton choices are independent for both the agent (π° and $\pi^{\bullet}$) and nature (θ° and $\theta^{\bullet}$), we can rewrite observation-based stickiness as:

$$
\begin{aligned}
= \sup_{\pi \in \Pi} \inf_{\theta^{mix} \in \Theta^{mix}} \Big\{ \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) \\
& + (95 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p) \\
& + (20 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a)) \cdot \theta^{det,\circ}(q) \\
& - 5 \cdot \pi^{\circ}(a) \cdot \theta^{det,\circ}(p) \cdot \theta^{det,\circ}(q) \\
& - 10 \cdot \pi^{\circ}(b) \cdot \theta^{det,\circ}(q)^2 \Big) \Big\} \\
+ \sup_{\pi \in \Pi} \inf_{\theta^{mix} \in \Theta^{mix}} \Big\{ \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 50 \cdot \pi^{\bullet}(a) \\
& + 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet}(p) \\
& + (200 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\bullet}(q) \\
& - 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet}(p) \cdot \theta^{det,\bullet}(q) \\
& - 100 \cdot \pi^{\bullet}(b) \cdot \theta^{det,\bullet}(q)^2 \Big) \Big\}.
\end{aligned}
$$

And the optimal value for the zero stickiness model:

$$
\begin{aligned}
V_{\text{fh}}^{*,M_3} = \sup_{\pi \in \Pi} \inf_{\theta^{mix} \in \Theta^{mix}} \Big\{ \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 45 \cdot \pi^{\circ}(b) + 5 \cdot \pi^{\circ}(a) + 50 \cdot \pi^{\bullet}(a) \\
& + (90 \cdot \pi^{\circ}(a) - 45 \cdot \pi^{\circ}(b)) \cdot \theta^{det,\circ}(p) \\
& + (10 \cdot \pi^{\circ}(b) - 5 \cdot \pi^{\circ}(a)) \cdot \theta^{det,\circ}(q) \\
& + 5 \cdot \pi^{\circ}(a) \cdot \theta^{det,\circ⁂}(p) \\
& + 10 \cdot \pi^{\circ}(b) \cdot \theta^{det,\circ⁂}(q) \\
& - 5 \cdot \pi^{\circ}(a) \cdot \theta^{det,\circ}(q) \cdot \theta^{det,\circ⁂}(p) \\
& - 10 \cdot \pi^{\circ}(b) \cdot \theta^{det,\circ}(q) \cdot \theta^{det,\circ⁂}(q) \\
& + (100 \cdot \pi^{\bullet}(b) - 50 \cdot \pi^{\bullet}(a)) \cdot \theta^{det,\bullet}(q) \\
& + 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet⁂}(p) \\
& + 100 \cdot \pi^{\bullet}(b) \cdot \theta^{det,\bullet⁂}(q) \\
& - 50 \cdot \pi^{\bullet}(a) \cdot \theta^{det,\bullet}(q) \cdot \theta^{det,\bullet⁂}(p) \\
& - 100 \cdot \pi^{\bullet}(b) \cdot \theta^{det,\bullet}(q) \cdot \theta^{det,\bullet⁂}(q) \Big) \Big\}.
\end{aligned}
$$

As histories ○○(○)(⁘) and ○●(○)(⁘), and ○○(○) and ○●(○) are mutually exclusive and the related non-singleton choices are independent for both the agent (π° and $\pi^{\bullet}$) and nature (θ°, $\theta^{\bullet}$, $\theta^{\circ\,⁘}$, and $\theta^{\bullet\,⁘}$), we can rewrite zero stickiness as:

$$
\begin{aligned}
= \sup_{\pi\in\Pi} \inf_{\theta^{mix}\in\Theta^{mix}} \Big\{ \sum_{\theta^{det}\in\Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 45\cdot\pi^{\circ}(b) + 5\cdot\pi^{\circ}(a) \\
& + (90\cdot\pi^{\circ}(a) - 45\cdot\pi^{\circ}(b))\cdot\theta^{det,\circ}(p) \\
& + (10\cdot\pi^{\circ}(b) - 5\cdot\pi^{\circ}(a))\cdot\theta^{det,\circ}(q) \\
& + 5\cdot\pi^{\circ}(a)\cdot\theta^{det,\circ⁘}(p) \\
& + 10\cdot\pi^{\circ}(b)\cdot\theta^{det,\circ⁘}(q) \\
& - 5\cdot\pi^{\circ}(a)\cdot\theta^{det,\circ}(q)\cdot\theta^{det,\circ⁘}(p) \\
& - 10\cdot\pi^{\circ}(b)\cdot\theta^{det,\circ}(q)\cdot\theta^{det,\circ⁘}(q)\Big)\Big\} \\
+ \sup_{\pi\in\Pi} \inf_{\theta^{mix}\in\Theta^{mix}} \Big\{ \sum_{\theta^{det}\in\Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 50\cdot\pi^{\bullet}(a) \\
& + (100\cdot\pi^{\bullet}(b) - 50\cdot\pi^{\bullet}(a))\cdot\theta^{det,\bullet}(q) \\
& + 50\cdot\pi^{\bullet}(a)\cdot\theta^{det,\bullet⁘}(p) \\
& + 100\cdot\pi^{\bullet}(b)\cdot\theta^{det,\bullet⁘}(q) \\
& - 50\cdot\pi^{\bullet}(a)\cdot\theta^{det,\bullet}(q)\cdot\theta^{det,\bullet⁘}(p) \\
& - 100\cdot\pi^{\bullet}(b)\cdot\theta^{det,\bullet}(q)\cdot\theta^{det,\bullet⁘}(q)\Big)\Big\}.
\end{aligned}
$$

We can further simplify it because p and q are independent:

$$
\begin{aligned}
= \sup_{\pi\in\Pi} \inf_{\theta^{mix}\in\Theta^{mix}} \Big\{ \sum_{\theta^{det}\in\Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 45\cdot\pi^{\circ}(b) + 5\cdot\pi^{\circ}(a) \\
& + (90\cdot\pi^{\circ}(a) - 45\cdot\pi^{\circ}(b))\cdot\theta^{det,\circ}(p) \\
& + (10\cdot\pi^{\circ}(b) - 5\cdot\pi^{\circ}(a))\cdot\theta^{det,\circ}(q) \\
& + (5\cdot\pi^{\circ}(a) - 5\cdot\pi^{\circ}(a)\cdot\theta^{det,\circ}(q))\cdot \inf_{\theta^{mix}\in\Theta^{mix}} \Big\{ \sum_{\theta^{det}\in\Theta^{det}} \theta^{mix}(\theta^{det})\cdot\theta^{det,\circ⁘}(p)\Big\} \\
& + (10\cdot\pi^{\circ}(b) - 10\cdot\pi^{\circ}(b)\cdot\theta^{det,\circ}(q))\cdot \inf_{\theta^{mix}\in\Theta^{mix}} \Big\{ \sum_{\theta^{det}\in\Theta^{det}} \theta^{mix}(\theta^{det})\cdot\theta^{det,\circ⁘}(q)\Big\}\Big)\Big\} \\
+ \sup_{\pi\in\Pi} \inf_{\theta^{mix}\in\Theta^{mix}} \Big\{ \sum_{\theta^{det}\in\Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \Big(& 50\cdot\pi^{\bullet}(a) \\
& + (100\cdot\pi^{\bullet}(b) - 50\cdot\pi^{\bullet}(a))\cdot\theta^{det,\bullet}(q) \\
& + (50\cdot\pi^{\bullet}(a) - 50\cdot\pi^{\bullet}(a)\cdot\theta^{det,\bullet}(q))\cdot \inf_{\theta^{mix}\in\Theta^{mix}} \Big\{ \sum_{\theta^{det}\in\Theta^{det}} \theta^{mix}(\theta^{det})\cdot\theta^{det,\bullet⁘}(p)\Big\} \\
& + (100\cdot\pi^{\bullet}(b) - 100\cdot\pi^{\bullet}(b)\cdot\theta^{det,\bullet}(q))\cdot \inf_{\theta^{mix}\in\Theta^{mix}} \Big\{ \sum_{\theta^{det}\in\Theta^{det}} \theta^{mix}(\theta^{det})\cdot\theta^{det,\bullet⁘}(q)\Big\}\Big)\Big\}.
\end{aligned}
$$

Table 4 displays the computed optimal values and policies, showing the differences between the full, observation-based, and zero stickiness assumptions.

Underlying POSGs

The POSG of the full stickiness interpretation of the RPOMDP in Figure 8 displays the same structural difference with the observation-based stickiness and zero stickiness interpretations as in Figure 4. The variable assignment chosen at the first nature state $\langle s_1, \{\}, \bot\rangle$ sticks in the full stickiness RPOMDP but not in the observation-based one zero stickiness ones.

The difference between the observation-based stickiness and zero stickiness POSGs occurs at a later stage, namely at nature states $\langle s4, \{\}, \bot\rangle$ and $\langle s5, \{\}, \bot\rangle$. After one of these states, the observation-based stickiness model follows the same structure as the full stickiness model, leading to infinitely many agent states with different variable restrictions. The POSG of the zero stickiness interpretation continues with infinitely many transitions going to the same agent states $\langle s_6, \{\}\rangle, \langle s_7, \{\}\rangle, \langle s_8, \{\}\rangle$, and $\langle s_9, \{\}\rangle$ with the totally undefined variable restriction.

	Full stickiness	Observation-based stickiness	Zero stickiness
Optimal value	$74\frac{11}{390}$	$71\frac{9}{10}$	$70\frac{295}{348}$
Optimal agent policy	$\bigcirc\bigcirc\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot&\cdot\end{smallmatrix}\right) \mapsto \{a \mapsto \frac{17}{117}, b \mapsto \frac{100}{117}\}$, $\bigcirc\bullet\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot&\cdot\end{smallmatrix}\right) \mapsto \{a \mapsto \frac{643}{1170}, b \mapsto \frac{527}{1170}\}$	$\bigcirc\bigcirc\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot&\cdot\end{smallmatrix}\right) \mapsto \{a \mapsto \frac{10}{31}, b \mapsto \frac{21}{31}\}$ $\bigcirc\bullet\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot&\cdot\end{smallmatrix}\right) \mapsto \{a \mapsto \frac{20}{31}, b \mapsto \frac{11}{31}\}$	$\bigcirc\bigcirc\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot&\cdot\end{smallmatrix}\right) \mapsto \{a \mapsto \frac{1}{3}, b \mapsto \frac{2}{3}\}$ $\bigcirc\bullet\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot&\cdot\end{smallmatrix}\right) \mapsto \{a \mapsto \frac{18}{29}, b \mapsto \frac{11}{29}\}$
Optimal nature policy	$\bigcirc \mapsto \{\{p \mapsto 0.1, q \mapsto 0.1\} \mapsto \frac{17}{24}$, $\{p \mapsto 0.9, q \mapsto 0.1\} \mapsto \frac{3}{104}$, $\{p \mapsto 0.9, q \mapsto 0.9\} \mapsto \frac{41}{156}\}$	$\bigcirc\bigcirc\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot\end{smallmatrix}\right) \mapsto \{$ $\{p \mapsto 0.1, q \mapsto 0.1\} \mapsto \frac{1663}{2232}$, $\{p \mapsto 0.9, q \mapsto 0.1\} \mapsto \frac{569}{2232}\}$ $\bigcirc\bullet\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot\end{smallmatrix}\right) \mapsto \{$ $\{p \mapsto 0.1, q \mapsto 0.1\} \mapsto \frac{187}{248}$, $\{p \mapsto 0.1, q \mapsto 0.9\} \mapsto \frac{61}{248}\}$	$\bigcirc\bigcirc\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot\end{smallmatrix}\right) \mapsto \{$ $\{p \mapsto 0.1, q \mapsto 0.1\} \mapsto \frac{1591}{2160}$, $\{p \mapsto 0.9, q \mapsto 0.1\} \mapsto \frac{569}{2160}\}$ $\bigcirc\bullet\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot\end{smallmatrix}\right) \mapsto \{$ $\{p \mapsto \_, q \mapsto 0.1\} \mapsto \frac{171}{232}$, $\{p \mapsto \_, q \mapsto 0.9\} \mapsto \frac{61}{232}\}$ $\bigcirc\bigcirc\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot&\cdot\end{smallmatrix}\right) \mapsto \{$ $\{p \mapsto 0.1, q \mapsto 0.1\} \mapsto 1\}$ $\bigcirc\bullet\!\!\left(\begin{smallmatrix}\cdot&\cdot\\\cdot&\cdot\end{smallmatrix}\right) \mapsto \{$ $\{p \mapsto 0.1, q \mapsto 0.1\} \mapsto 1\}$

Table 4: Optimal values and policies for the full stickiness, observation-based stickiness, and zero stickiness interpretations of the RPOMDP in Figure 8.

D.3 Order of Play Matters

We first revisit the RPOMDP in Figure 3 and show how we computed the optimal values to show that order of play matters in RPOMDPs.

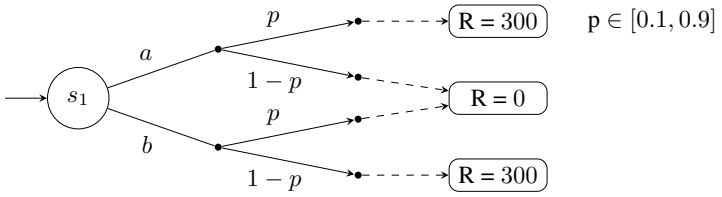


Figure 3: An RPOMDP where agent first and nature first semantics do not coincide in their optimal value.

For $\pi \in \Pi$, we write $\pi^{\bigcirc} = \pi(\bigcirc)$. Similarly, for $\theta \in \Theta$ of the agent first RPOMDP, we write $\theta^a = \theta(\langle \bigcirc, a \rangle)$ and $\theta^b = \theta(\langle \bigcirc, b \rangle)$, and for $\theta \in \Theta$ of the nature first RPOMDP, we write $\theta^{\bigcirc} = \theta(\bigcirc)$. Using this notation, we can construct the value functions for the agent first and nature interpretations of the RPOMDP in Figure 3. We construct these value functions following the same approach as for the value functions of the RPOMDP in Figure 2, see Appendix D.1 and Figure 7. The value function for the agent first RPOMDP M_1 is:

$$V_{\text{fh}}^{M_1}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \Big(\pi^{\bigcirc}(a) \cdot \theta^{det,a}(p) \cdot 300 + \pi^{\bigcirc}(b) \cdot (1 - \theta^{det,b}(p)) \cdot 300 \Big).$$

And the value function for the nature first RPOMDP M_2:

$$V_{\text{fh}}^{M_2}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \Big(\pi^{\bigcirc}(a) \cdot \theta^{det,\bigcirc}(p) \cdot 300 + \pi^{\bigcirc}(b) \cdot (1 - \theta^{det,\bigcirc}(p)) \cdot 300 \Big).$$

Both these functions are linear in the deterministic nature policies. As we again have a convex uncertainty set, we can follow the same steps as for Proposition 3 and restrict the search for the optimal nature policy to the set of deterministic nature policies.

Using the above functions, we can compute the optimal value for the agent first model as follows:

$$V_{\text{fh}}^{*,M_1} = \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} \Big\{ \pi^{\bigcirc}(a) \cdot \theta^{det,a}(p) \cdot 300 + \pi^{\bigcirc}(b) \cdot (1 - \theta^{det,b}(p)) \cdot 300 \Big\}.$$

And the optimal value for the nature first model:

$$V_{\text{fh}}^{*,M_2} = \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} \Big\{ \pi^{\bigcirc}(a) \cdot \theta^{det,\bigcirc}(p) \cdot 300 + \pi^{\bigcirc}(b) \cdot (1 - \theta^{det,\bigcirc}(p)) \cdot 300 \Big\}.$$

Table 5 displays the computed optimal values and policies, showing the differences between the agent and nature first assumptions. An underscore indicates that the choice at this history does not influence the optimal value of the RPOMDP.

	Agent first	Nature first
Optimal value	30	150
Optimal agent policy	$\bigcirc \mapsto \_$	$\bigcirc \mapsto \{a \mapsto 0.5, b \mapsto 0.5\}$
Optimal nature policy	$\langle \bigcirc, a \rangle \mapsto \{p \mapsto 0.1\}$, $\langle \bigcirc, b \rangle \mapsto \{p \mapsto 0.9\}$	$\bigcirc \mapsto \{p \mapsto 0.5\}$

Table 5: Optimal values and policies for the agent first and nature first interpretations of the RPOMDP in Figure 3.

Underlying POSGs

Figure 5 (restated below) depicts the agent first and nature first POSGs of the RPOMDP in Figure 3. We briefly discuss the structural differences. As the agent has a finite choice of actions, this will always lead to a finite split in the POSG. Nature's number of choices depends on the variable restrictions in the nature state. In this simple model, we only have unrestricted, and hence infinite, nature choices. The structural difference between the two POSGs is caused entirely by the order of play.

When the agent chooses first, the POSG has a finite number of states. The infinite choice in the nature states $\langle s_1, \{\}, a \rangle$ and $\langle s_1, \{\}, b \rangle$ all lead to the same reward states, just with different probabilities determined by the chosen variable assignment.

When nature chooses first, we get an infinite number of agent states after nature's infinite choice in nature state $\langle s_1, \{\} \rangle$, as the chosen variable assignment needs to be recorded for the transition after the agent's choice. The number of states in the nature first model is hence infinite. Each resulting agent state only has a finite choice leading to the same reward states.

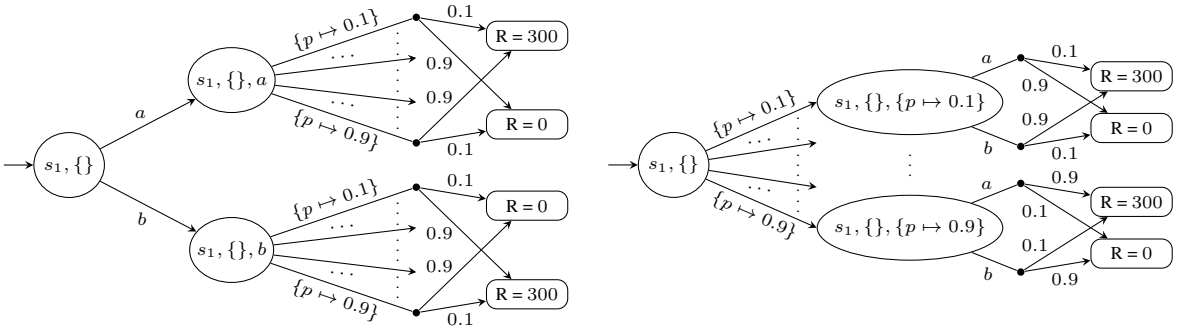


Figure 5: Agent first (left) and nature first (right) POSGs of the RPOMDP in Figure 3.

D.4 a-Rectangularity

Next, we look at an a-rectangular RPOMDP to show that order of play still matters under a form of rectangularity and is not only a concern in non-rectangular RPOMDPs. Consider the RPOMDP in Figure 9. We interpret this RPOMDP with full stickiness semantics.

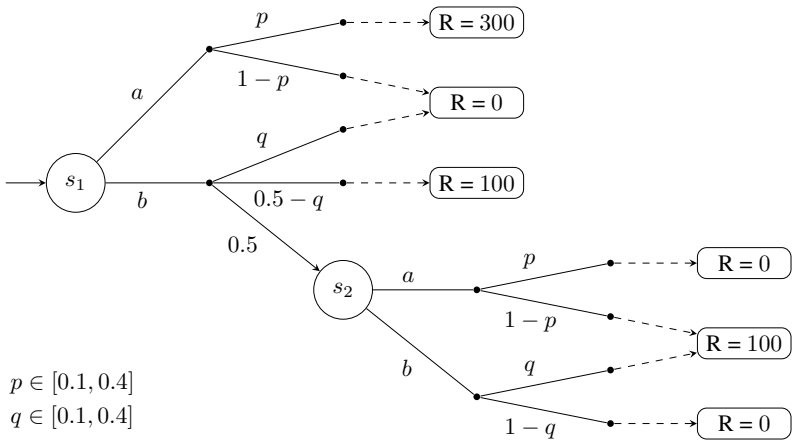


Figure 9: An a-rectangular RPOMDP where agent first and nature first semantics do not coincide in their optimal value.

For $\pi \in \Pi$, we write $\pi^{\circ} = \pi(\bigcirc)$ and $\pi^{\circ\circ} = \pi(\bigcirc\bigcirc)$. Similarly, for $\theta \in \Theta$ of the agent first RPOMDP, we write $\theta^a = \theta(\langle \bigcirc, a \rangle)$ and $\theta^b = \theta(\langle \bigcirc, b \rangle)$, and for $\theta \in \Theta$ of the nature first RPOMDP, we write $\theta^{\circ} = \theta(\bigcirc)$. Using this notation, we can construct the value functions for the agent first and nature interpretations of the RPOMDP in Figure 9. We construct these value functions following the same approach as for the value functions of the RPOMDP in Figure 2, see Appendix D.1 and Figure 7. The value function for the agent first RPOMDP M_1 is:

$$V_{\text{fh}}^{M_1}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \Big(\pi^{\circ}(a) \cdot \theta^{det,a}(p) \cdot 300 + \pi^{\circ}(b) \cdot \big((0.5 - \theta^{det,b}(q)) \cdot 100 + 0.5 \cdot (\pi^{\circ\circ}(a) \cdot (1 - \theta^{det,b}(p)) \cdot 100 + \pi^{\circ\circ}(b) \cdot \theta^{det,b}(q) \cdot 100) \big) \Big).$$

And the value function for the nature first RPOMDP M_2:

$$V_{\text{fh}}^{M_1}(\pi, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \Big(\pi^{\circ}(a) \cdot \theta^{det\circ}(p) \cdot 300 + \pi^{\circ}(b) \cdot \big((0.5 - \theta^{det\circ}(q)) \cdot 100 + 0.5 \cdot (\pi^{\circ\circ}(a) \cdot (1 - \theta^{det\circ}(p)) \cdot 100 + \pi^{\circ\circ}(b) \cdot \theta^{det\circ}(q) \cdot 100) \big) \Big).$$

Both these functions are linear in the deterministic nature policies. As we again have a convex uncertainty set, we can follow the same steps as for Proposition 3 and restrict the search for the optimal nature policy to the set of deterministic nature policies. Using the above functions, we can compute the optimal value for the agent first model as follows:

$$V_{\text{fh}}^{*,M_1} = \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} \Big\{ \pi^{\circ}(a) \cdot \theta^{det,a}(p) \cdot 300 + \pi^{\circ}(b) \cdot \big((0.5 - \theta^{det,b}(q)) \cdot 100 + 0.5 \cdot (\pi^{\circ\circ}(a) \cdot (1 - \theta^{det,b}(p)) \cdot 100 + \pi^{\circ\circ}(b) \cdot \theta^{det,b}(q) \cdot 100) \big) \Big\}.$$

And the optimal value for the nature first RPOMDP M_2:

$$V_{\text{fh}}^{*,M_2} = \sup_{\pi \in \Pi} \inf_{\theta^{det} \in \Theta^{det}} \Big\{ \pi^{\circ}(a) \cdot \theta^{det,\circ}(p) \cdot 300 + \pi^{\circ}(b) \cdot \big((0.5 - \theta^{det,\circ}(q)) \cdot 100 + 0.5 \cdot (\pi^{\circ\circ}(a) \cdot (1 - \theta^{det,\circ}(p)) \cdot 100 + \pi^{\circ\circ}(b) \cdot \theta^{det,\circ}(q) \cdot 100) \big) \Big\}.$$

Table 6 displays the computed optimal values and policies, showing the differences between the agent and nature first assumptions. An underscore indicates that the value assigned to this variable does not influence the optimal value of the RPOMDP.

Underlying POSGs

Figure 10 depicts the agent first and nature first POSGs of the RPOMDP in Figure 9. The structural difference between these POSGs, like between the POSGs in Figure 5, is caused by the different points of infinite branching.

	Agent first	Nature first
Optimal value	40	$51\frac{3}{7}$
Optimal agent policy	$\bigcirc \mapsto \{a \mapsto 0, b \mapsto 1\}$, $\bigcirc\bigcirc \mapsto \{a \mapsto 1, b \mapsto 0\}$	$\bigcirc \mapsto \{a \mapsto \frac{1}{7}, b \mapsto \frac{6}{7}\}$, $\bigcirc\bigcirc \mapsto \{a \mapsto 1, b \mapsto 0\}$
Optimal nature policy	$\langle \bigcirc, a \rangle \mapsto \{p \mapsto 0.1, q \mapsto \_\}$, $\langle \bigcirc, b \rangle \mapsto \{p \mapsto 0.4, q \mapsto 0.4\}$	$\bigcirc \mapsto \{p \mapsto \frac{6}{35}, q \mapsto 0.4\}$

Table 6: Optimal values and agent policies for the agent first and nature first interpretations of the RPOMDP in Figure 9.

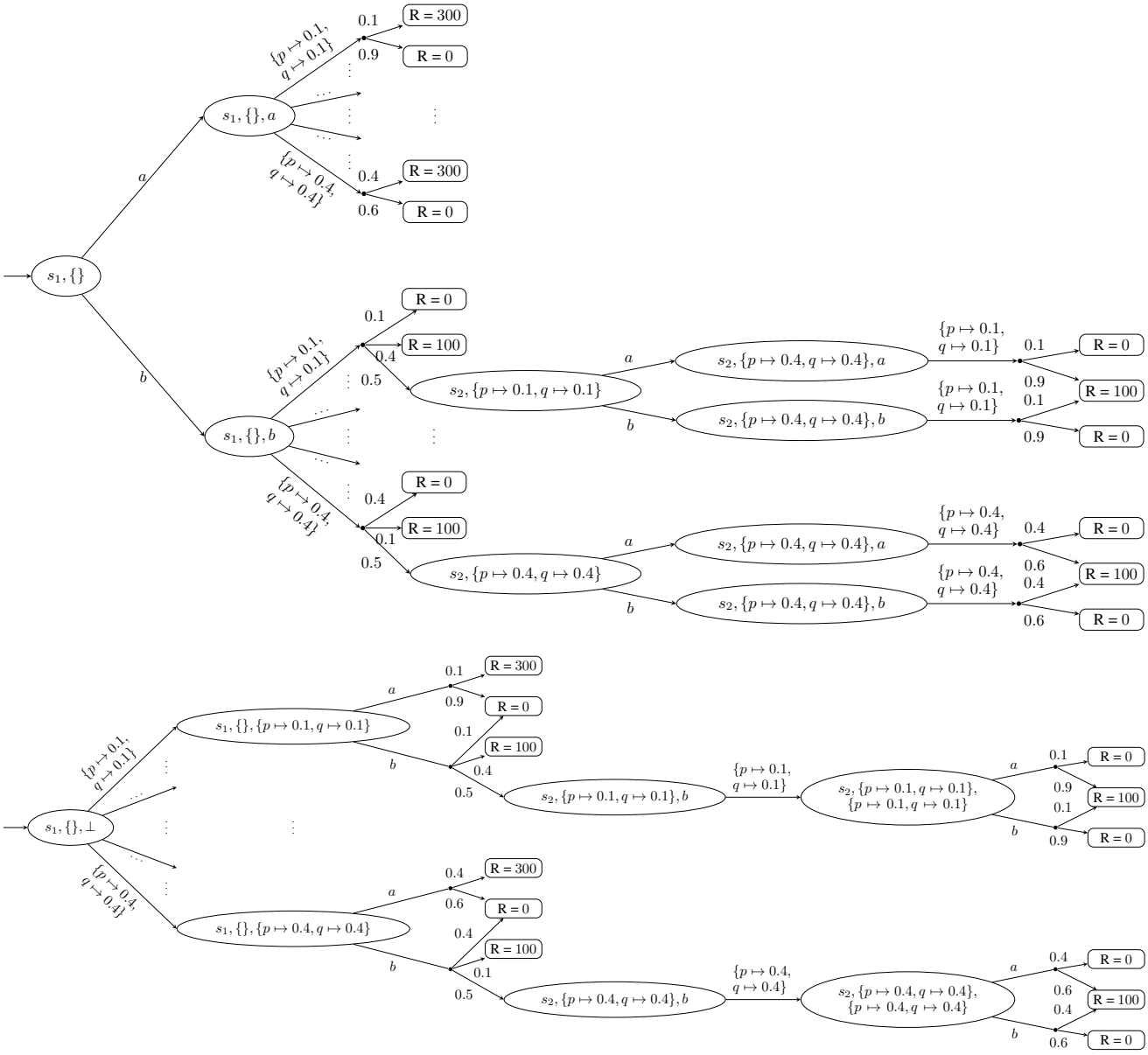


Figure 10: Agent first (top) and nature first (bottom) POSGs of the RPOMDP in Figure 9.

E Nature First Semantics

Throughout the main paper and the appendix, the definitions and proofs are all written with the agent first order of play. This appendix discusses the changes required to achieve the same results with the nature first semantics.

Policies in the RPOMDP. When nature moves first, nature receives the agent's action after choosing its own action. Therefore, nature policies in nature first RPOMDPs are of the following types:

$$\begin{aligned} \text{Stochastic:} \quad & \theta \colon H^{\mathfrak{n},M} \to \Delta(\boldsymbol{U}), \\ \text{Deterministic:} \quad & \theta^{det} \colon H^{\mathfrak{n},M} \to \boldsymbol{U}, \\ \text{Mixed:} \quad & \theta^{mix} \in \Delta(H^{\mathfrak{n},M} \to \boldsymbol{U}). \end{aligned}$$

The agent policies do not change, as we assume the agent still cannot observe the variable assignments nature chooses.

Nature first POSG. Given a nature first RPOMDP, we define its POSG as follows.

Definition 18 (Equivalent nature first POSG)**.** *Given a robust POMDP $\langle S, A, \boldsymbol{T}, R, Z_\bullet^{\mathfrak{a}}, Z_\bullet^{\mathfrak{n}}, Z_\circ, O_\bullet^{\mathfrak{a}}, O_\bullet^{\mathfrak{n}}, O_\circ \rangle$, we define the POSG where nature chooses first as a tuple $\langle \mathcal{S}^{\mathfrak{a}}, \mathcal{S}^{\mathfrak{n}}, \mathcal{A}^{\mathfrak{a}}, \mathcal{A}^{\mathfrak{n}}, \mathcal{T}, \mathcal{R}, \mathcal{Z}^{\mathfrak{a}}, \mathcal{Z}^{\mathfrak{n}}, \mathcal{O}^{\mathfrak{a}}, \mathcal{O}^{\mathfrak{n}} \rangle$, where $\mathcal{S}^{\mathfrak{n}}, \mathcal{A}^{\mathfrak{a}}, \mathcal{A}^{\mathfrak{n}}, \mathcal{Z}^{\mathfrak{a}}$, and $\mathcal{Z}^{\mathfrak{n}}$ remain the same as in Definition 5. The agent's state-space is given by $\mathcal{S}^{\mathfrak{a}} = S \times \boldsymbol{U}^{\hookrightarrow} \times \boldsymbol{U}$, and the transition, reward, and observation functions are defined as follows:*

- $\mathcal{T}^{\mathfrak{a}} \colon \mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \to \Delta(\mathcal{S}^{\mathfrak{n}})$*, by* $\mathcal{T}^{\mathfrak{a}}(\langle s, u^{\hookrightarrow}, u \rangle, a, \langle s', \mathsf{upd}(u^{\hookrightarrow}, u, O_\bullet^{\mathfrak{n}}(s), O_\circ(s), a), a \rangle) = \boldsymbol{T}(u)(s, a, s')$.
- $\mathcal{T}^{\mathfrak{n}} \colon \mathcal{S}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}} \to \mathcal{S}^{\mathfrak{a}}$*, by* $\mathcal{T}^{\mathfrak{n}}(\langle s, u^{\hookrightarrow}, a \rangle, u, \langle s, u^{\hookrightarrow}, u \rangle) = \begin{cases} 1 & \text{if } u \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookrightarrow}), \\ 0 & \text{otherwise.} \end{cases}$
- $\mathcal{R} \colon \mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \to \mathbb{R}$ *by* $\mathcal{R}(\langle s, u^{\hookrightarrow}, u \rangle, a) = R(s, a)$.
- $\mathcal{O}^{\mathfrak{a}} \colon (\mathcal{S}^{\mathfrak{a}} \cup \mathcal{S}^{\mathfrak{n}}) \to \mathcal{Z}^{\mathfrak{a}}$ *by* $\mathcal{O}^{\mathfrak{a}}(s) = \begin{cases} \langle O_\bullet^{\mathfrak{a}}(s'), O_\circ(s') \rangle & \text{if } s = \langle s', u^{\hookrightarrow}, u \rangle \in \mathcal{S}^{\mathfrak{a}}, \\ \langle O_\bullet^{\mathfrak{a}}(s'), O_\circ(s') \rangle & \text{if } s = \langle s', u^{\hookrightarrow}, a \rangle \in \mathcal{S}^{\mathfrak{n}}. \end{cases}$
- $\mathcal{O}^{\mathfrak{n}} \colon (\mathcal{S}^{\mathfrak{a}} \cup \mathcal{S}^{\mathfrak{n}}) \to \mathcal{Z}^{\mathfrak{n}}$ *by*
  $\mathcal{O}^{\mathfrak{n}}(s) = \begin{cases} \langle O_\bullet^{\mathfrak{n}}(s'), O_\circ(s'), \bot \rangle & \text{if } s = \langle s', u^{\hookrightarrow}, u \rangle \in \mathcal{S}^{\mathfrak{a}}, \\ \langle O_\bullet^{\mathfrak{n}}(s'), O_\circ(s'), a \rangle & \text{if } s = \langle s', u^{\hookrightarrow}, a \rangle \in \mathcal{S}^{\mathfrak{n}}. \end{cases}$

The a observed in a $\mathcal{S}^{\mathfrak{n}}$ state corresponds to the previously chosen a. So, the action that nature observes in a nature state $\langle s', u^{\hookrightarrow}, a \rangle \in \mathcal{S}^{\mathfrak{n}}$ is the action $a \in A$ that was taken to reach the current state $s' \in S$ of the RPOMDP, not the action the agent will take from the current state. This game starts in a $\mathcal{S}^{\mathfrak{n}}$ state consisting of the initial state $s_I \in S$ in the RPOMDP, the totally undefined variable assignment $u^{\perp} \in \boldsymbol{U}^{\hookrightarrow}$, and a placeholder for the action $\bot$.

Paths and histories. When reasoning with the nature first semantics, the order of the paths in the POSG changes:

$$\text{Paths}^G : (\mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \times \mathcal{S}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}})^* \times \mathcal{S}^{\mathfrak{a}} \Longrightarrow (\mathcal{S}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}} \times \mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}})^* \times \mathcal{S}^{\mathfrak{n}}.$$

As a result, the histories similarly change:

$$\begin{aligned} H^G &: (\mathcal{Z}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}})^* \times \mathcal{Z}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{n}} \Longrightarrow (\mathcal{Z}^{\mathfrak{n}} \times \mathcal{Z}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{n}} \times \mathcal{Z}^{\mathfrak{n}} \times \mathcal{Z}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}})^* \times \mathcal{Z}^{\mathfrak{n}} \times \mathcal{Z}^{\mathfrak{a}}. \\ H^{\mathfrak{a},G} &: (\mathcal{Z}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{a}})^* \times \mathcal{Z}^{\mathfrak{a}} \Longrightarrow (\mathcal{Z}^{\mathfrak{a}} \times \mathcal{Z}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}})^* \times \mathcal{Z}^{\mathfrak{a}}. \\ H^{\mathfrak{n},G} &: (\mathcal{Z}^{\mathfrak{n}} \times \mathcal{Z}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}})^* \times \mathcal{Z}^{\mathfrak{n}} \Longrightarrow (\mathcal{Z}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}} \times \mathcal{Z}^{\mathfrak{n}})^* \times \mathcal{Z}^{\mathfrak{n}}. \end{aligned}$$

Policies in the POSG As the order of the paths and historis changed, the policy types also change. The agent now observes an extra state, instead of nature. Note that this extra observation contains no extra information for the agent, while it did for nature in the agent-first semantics.

$$\begin{aligned} \text{Stochastic:} \quad & \pi \colon H^{\mathfrak{a},G} \to \Delta(\mathcal{A}^{\mathfrak{a}}) && \Longrightarrow && \pi \colon H^{\mathfrak{a},G} \times \mathcal{Z}^{\mathfrak{a}} \to \Delta(\mathcal{A}^{\mathfrak{a}}), \\ \text{Deterministic:} \quad & \pi^{det} \colon H^{\mathfrak{a},G} \to \mathcal{A}^{\mathfrak{a}} && \Longrightarrow && \pi^{det} \colon H^{\mathfrak{a},G} \times \mathcal{Z}^{\mathfrak{a}} \to \mathcal{A}^{\mathfrak{a}}, \\ \text{Mixed:} \quad & \pi^{mix} \in \Delta(H^{\mathfrak{a},G} \to \mathcal{A}^{\mathfrak{a}}) && \Longrightarrow && \pi^{mix} \in \Delta(H^{\mathfrak{a},G} \times \mathcal{Z}^{\mathfrak{a}} \to \mathcal{A}^{\mathfrak{a}}), \end{aligned}$$

$$\begin{aligned} \text{Stochastic:} \quad & \theta \colon H^{\mathfrak{n},G} \times \mathcal{Z}^{\mathfrak{n}} \to \Delta(\mathcal{A}^{\mathfrak{n}}), && \Longrightarrow && \theta \colon H^{\mathfrak{n},G} \to \Delta(\mathcal{A}^{\mathfrak{n}}) \\ \text{Deterministic:} \quad & \theta^{det} \colon H^{\mathfrak{n},G} \times \mathcal{Z}^{\mathfrak{n}} \to \mathcal{A}^{\mathfrak{n}}, && \Longrightarrow && \theta^{det} \colon H^{\mathfrak{n},G} \to \mathcal{A}^{\mathfrak{n}} \\ \text{Mixed:} \quad & \theta^{mix} \in \Delta(H^{\mathfrak{n},G} \times \mathcal{Z}^{\mathfrak{n}} \to \mathcal{A}^{\mathfrak{n}}) && \Longrightarrow && \theta^{mix} \in \Delta(H^{\mathfrak{n},G} \to \mathcal{A}^{\mathfrak{n}}). \end{aligned}$$

Additional adaptations. At the end of Appendices A, F and G, the adjustments required for the definitions or proofs to work with the nature first semantics are briefly discussed.

F Equivalent Values

Given the stickiness and order of play, we show that the value of an RPOMDP and its POSG are equivalent (Theorem 2). To do so, we construct a bijection between the sets of paths of the two models. We use this bijection to subsequently construct new bijections between the sets of histories and policies and finally conclude that the values are equivalent. For convenience, we repeat the proposition from the main text and then split it into several lemmas.

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

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

Lemma 2 (Bijection between paths)**.** *Let M be an RPOMDP, and G the POSG of M. There exists a bijection $f\colon \mathit{Paths}^M \to \mathit{Paths}^G$.*

Proof. Let $\text{Paths}^{M,\ltimes} \subseteq (S \times A \times \boldsymbol{U})^* \times S^?$ with $? \in \{0,1\}$ be the set of all path segments in the RPOMDP, and let $\text{Paths}^{G,\ltimes} \subseteq (\mathcal{S}^{\mathtt{a}} \times \mathcal{A}^{\mathtt{a}} \times \mathcal{S}^{\mathtt{n}} \times \mathcal{A}^{\mathtt{n}})^* \times (\mathcal{S}^{\mathtt{a}})^?$ with $? \in \{0,1\}$ be the set of all path segments in the POSG. With path segment we mean that the path can starts at any time steps $t \in \mathbb{N}$ and can end at any time step $t' \in \mathbb{N}, t <= t'$. The optional last state is only used for path segments until the horizon. Note that $\text{Paths}^M \subseteq \text{Paths}^{M,\ltimes}$ and $\text{Paths}^G \subseteq \text{Paths}^{G,\ltimes}$. Let $\tau^M = \langle s_0, a_0, u_0, s_1, \ldots, s_n\rangle \in \text{Paths}^M$ and $t \leq n$, then $\tau^M(t)$ indicates the t-th segment $\langle s_t, a_t, u_t\rangle$ of τ^M. Note that if $t = n$, the segment will only consist of the final state $\langle s_n\rangle$. Similarly, let $\tau^G = \langle s_0^{\mathtt{a}}, a_0^{\mathtt{a}}, s_0^{\mathtt{n}}, a_0^{\mathtt{n}}, s_1^{\mathtt{a}}, \ldots, s_n^{\mathtt{a}}\rangle = \langle\langle s_0, u^{\hookrightarrow}\rangle, a_0, \langle s_0, u^{\hookrightarrow}, a_0\rangle, u_0, \langle s_1, u_1^{\hookrightarrow}\rangle, \ldots, \langle s_n, u_n^{\hookrightarrow}\rangle\rangle \in \text{Paths}^G$ and $t \leq n$, then $\tau^G(t)$ indicates the t-th segment $\langle s_t^{\mathtt{a}}, a_t^{\mathtt{a}}, s_t^{\mathtt{n}}, a_t^{\mathtt{n}}\rangle = \langle\langle s_t, u_t^{\hookrightarrow}\rangle, a_t, \langle s_t, u_t^{\hookrightarrow}, a_t\rangle, u_t\rangle$ of τ^G. Note that if $t = n$, the segment will only consist of the final agent state $\langle s_n^{\mathtt{a}}\rangle = \langle\langle s_n, u_n^{\hookrightarrow}\rangle\rangle$.

Let $g\colon \text{Paths}^{M,\ltimes} \times \boldsymbol{U}^{\hookrightarrow} \hookrightarrow \text{Paths}^{G,\ltimes}$ defined by:

$$g(\langle s\rangle, u^{\hookrightarrow}) = \langle\langle s, u^{\hookrightarrow}\rangle\rangle.$$

$$g(\langle s, a, u\rangle, u^{\hookrightarrow}) = \begin{cases} \langle\langle s, u^{\hookrightarrow}\rangle, a, \langle s, u^{\hookrightarrow}, a\rangle, u\rangle & \text{if } u \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookrightarrow}),\\ \bot & \text{otherwise.}\end{cases}$$

$$g(\langle s, a, u\rangle \oplus \tau^{M\prime}, u^{\hookrightarrow}) = \begin{cases} g(\langle s, a, u\rangle, u^{\hookrightarrow}) \oplus g(\tau^{M\prime}, \mathsf{upd}(u^{\hookrightarrow}, u, O_{\bullet}^{\mathtt{n}}(s), O_{\circ}(s), a)) & \text{if } u \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookrightarrow}),\\ \bot & \text{otherwise.}\end{cases}$$

Let $f\colon \text{Paths}^M \to \text{Paths}^G$ defined by:

$$f(\langle s\rangle) = \langle\langle s, u^{\bot}\rangle\rangle.$$

$$f(\langle s, a, u\rangle) = \langle\langle s, u^{\bot}\rangle, a, \langle s, u^{\bot}, a\rangle, u\rangle.$$

$$f(\langle s, a, u\rangle \oplus \tau^{M\prime}) = f(\langle s, a, u\rangle) \oplus g(\tau^{M\prime}, \mathsf{upd}(u^{\bot}, u, O_{\bullet}^{\mathtt{n}}(s), O_{\circ}(s), a)).$$

Where $u^{\bot} \in \boldsymbol{U}^{\hookrightarrow}$ is the totally undefined function. Note that the results of f and g are in Paths^G and $\text{Paths}^{G,\ltimes}$ by construction. Also, note that any call to g that originated from a call in f will have a result by construction.

We show that f is a bijection, meaning f is injective and surjective. We first show f is injective, so we show that:

$$\forall \tau^{1,M}, \tau^{2,M} \in \text{Paths}^M . \tau^{1,M} \neq \tau^{2,M} \implies f(\tau^{1,M}) \neq f(\tau^{2,M}).$$

Given arbitrary $\tau^{1,M}, \tau^{2,M} \in \text{Paths}^M$, we distinguish between the paths with superscripts 1 and 2, respectively. Assume $\tau^{1,M} \neq \tau^{2,M}$. If $\tau^{1,M}$ and $\tau^{2,M}$ do not have the same horizon length, then neither do $f(\tau^{1,M})$ and $f(\tau^{2,M})$. Then trivially, $f(\tau^{1,M}) \neq f(\tau^{2,M})$.

Assume $\tau^{1,M}$ and $\tau^{2,M}$ have the same horizon length n. Then $\exists t \leq n$ where $\tau^{1,M}$ and $\tau^{2,M}$ deviate, so $\tau^{1,M}(t) \neq \tau^{2,M}(t)$. Let q be the smallest number where the paths deviate. So $\forall t < q . \tau^{1,M}(t) = \tau^{2,M}(t)$ and $\tau^{1,M}(q) \neq \tau^{2,M}(q)$. Assume $q < n$. Then we know $\langle s_q^1, a_q^1, u_q^1\rangle \neq \langle s_q^2, a_q^2, u_q^2\rangle$, which comes down to: $s_q^1 \neq s_q^2 \vee a_q^1 \neq a_q^2 \vee u_q^1 \neq u_q^2$.

$$\begin{aligned} f(\tau^{1,M}) &= f(\tau^{1,M}(1) \oplus \tau^{1,M\prime})\\ &= \tau^{1,G}(1) \oplus g(\tau^{1,M\prime}, \mathsf{upd}(u^{\bot}, u_0^1, O_{\bullet}^{\mathtt{n}}(s_0^1), O_{\circ}(s_0^1), a_0^1)).\end{aligned}$$

Unfold g until q:

$$= \bigoplus_{t=0}^{q} (\tau^{1,G}(t)) \oplus \langle \langle s_q^1, \mathsf{fix}(\tau_{0:q}^{1,M}) \rangle, a_q^1, \langle s_q^1, \mathsf{fix}(\tau_{0:q}^{1,M}), a_q^1 \rangle, u_q^1 \rangle \oplus g(\tau^{1,M\prime\prime}, \mathsf{fix}(\tau_{0:q+1}^{1,M})).$$

Since $s_q^1 \neq s_q^2 \vee a_q^1 \neq a_q^2 \vee u_q^1 \neq u_q^2$:

$$\neq \bigoplus_{t=0}^{q} (\tau^{1,G}(t)) \oplus \langle \langle s_q^2, \mathsf{fix}(\tau_{0:q}^{1,M}) \rangle, a_q^2, \langle s_q^2, \mathsf{fix}(\tau_{0:q}^{1,M}), a_q^2 \rangle, u_q^2 \rangle \oplus g(\tau^{2,M\prime\prime}, \mathsf{fix}(\tau_{0:q+1}^{1,M}))$$
$$= \bigoplus_{t=0}^{q} (\tau^{2,G}(t)) \oplus \langle \langle s_q^2, \mathsf{fix}(\tau_{0:q}^{2.M}) \rangle, a_q^2, \langle s_q^2, \mathsf{fix}(\tau_{0:q}^{2,M}), a_q^2 \rangle, u_q^2 \rangle \oplus g(\tau^{2,M\prime\prime}, \mathsf{fix}(\tau_{0:q+1}^{2,M})).$$

Fold g until 1:

$$= \tau^{2,G}(1) \oplus g(\tau^{2,M\prime}, \mathsf{upd}(u^{\perp}, u_0^2, O_{\bullet}^{\mathfrak{n}}(s_0^2), O_{\circ}(s_0^2), a_0^2))$$
$$= f(\tau^{2,M}(1) \oplus \tau^{2,M\prime})$$
$$= f(\tau^{2,M}).$$

If $q = n$, then the same result follows by removing everything after $\langle s_q^1, \mathsf{fix}(\tau_{0:q}^{1,M}) \rangle$, $\langle s_q^2, \mathsf{fix}(\tau_{0:q}^{1,M}) \rangle$, and $\langle s_q^2, \mathsf{fix}(\tau_{0:q}^{2,M}) \rangle$. We thus have that $f(\tau^{1,M}) \neq f(\tau^{2,M})$, so f is injective.

Next, we show that f is surjective, so we show that:

$$\forall \tau^G \in \text{Paths}^G, \exists \tau^M \in \text{Paths}^M . f(\tau^M) = \tau^G.$$

We show this holds by induction on the horizon length of the $\tau^G \in \text{Paths}^G$. We write the length of τ^G as $|\tau^G|$.

Assume $|\tau^G| = 0$. Then $\tau^G = \langle s_I, u^{\perp} \rangle$. We have that for $\langle s_I \rangle \in \text{Paths}^M$, $f(\langle s_I \rangle) = \langle \langle s_I, u^{\perp} \rangle \rangle = \tau^G$. So for paths of horizon length 0, f is surjective.

Now assume we know, given $q \in \mathbb{N}, q \geq 1$, that:

$$\forall \tau^G \in \text{Paths}^G . |\tau^G| = q - 1 \implies \exists \tau^M \in \text{Paths}^M : f(\tau^M) = \tau^G.$$

Take arbitrary $\tau^G \in \text{Paths}^G$ with horizon length $|\tau^G| = q$. Then we have $\tau^G = \tau_{0:q-1}^G \oplus \langle a_{q-1}, \langle s_{q-1}, u_{q-1}^{\hookrightarrow}, a_{q-1} \rangle, u_{q-1}, \langle s_q, u_q^{\hookrightarrow} \rangle \rangle$. Then $\tau_{0:q-1}^G \in \text{Paths}^G$ and $|\tau_{0:q-1}^G| = q - 1$. By assumption, we get that:

$$\exists \tau_{0:q-1}^M \in \text{Paths}^M, f(\tau_{0:q-1}^M) = \tau_{0:q-1}^G.$$

Let $\tau_{0:q-1}^M \in \text{Paths}^M$ such that $f(\tau_{0:q-1}^M) = \tau_{0:q-1}^G$. We then know that in τ^G:

$$\forall t < q . u_t^{\hookrightarrow} = \mathsf{fix}(\tau_{0:t}^M).$$

And, by Definition 5 and the definition of Paths^G, that:

$$u_q^{\hookrightarrow} = \mathsf{upd}(\mathsf{fix}(\tau_{0:q-1}^M), u_{q-1}, O_{\bullet}^{\mathfrak{n}}(s_{q-1}), O_{\circ}(s_{q-1}), a_{q-1}).$$

Let $\langle \langle s_{q-2}, \mathsf{fix}(\tau_{0:q-2}^M) \rangle, a_{q-2}, \langle s_{q-2}, \mathsf{fix}(\tau_{0:q-2}^M), a_{q-2} \rangle, u_{q-2}, \langle s_{q-1}, \mathsf{fix}(\tau_{0:q-1}^M) \rangle \rangle$ be the last two segments of $\tau_{0:q-1}^G$. Then by definition and injectivity of f, we know that the last two segments of $\tau_{0:q-1}^M$ are $\langle s_{q-2}, a_{q-2}, u_{q-2}, s_{q-1} \rangle$.

Now, by definition Definition 5 and the definition of Paths^G, we know that:

$$\tau^G = \tau_{0:q-1}^G \oplus \langle a_{q-1}, \langle s_{q-1}, u_{q-1}^{\hookrightarrow}, a_{q-1} \rangle, u_{q-1}, \langle s_q, u_q^{\hookrightarrow} \rangle \rangle \in \text{Paths}^G$$
$$\iff$$
$$\tau_{0:q-1}^G \in \text{Paths}^G \wedge \mathcal{T}^{\mathfrak{a}}(\langle s_{q-1}, u_{q-1}^{\hookrightarrow} \rangle, a_{q-1}, \langle s_{q-1}, u_{q-1}^{\hookrightarrow}, a_{q-1} \rangle) > 0$$
$$\wedge \mathcal{T}^{\mathfrak{n}}(\langle s_{q-1}, u_{q-1}^{\hookrightarrow}), a_{q-1} \rangle, u_{q-1}, \langle s_q, u_q^{\hookrightarrow} \rangle) > 0$$
$$\iff$$

$$\tau^G_{0:q-1} \in \text{Paths}^G \wedge u_{q-1} \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookrightarrow}_{q-1}) \wedge \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) > 0$$
$$\iff$$
$$\tau^G_{0:q-1} \in \text{Paths}^G \wedge u_{q-1} \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(\tau^M_{0:q-1})) \wedge \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) > 0.$$

So, since $\tau^G \in \text{Paths}^G$, we know $u_{q-1} \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(\tau^M_{0:q-1}))$ and $\boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) > 0$, which are the restrictions for $\tau^M = \tau^M_{0:q-1} \oplus \langle a_{q-1}, u_{q-1}, s_q \rangle \in \text{Paths}^M$ to hold.

$$\begin{aligned}
f(\tau^M) &= f(\tau^M_{0:q-1} \oplus \langle a_{q-1}, u_{q-1}, s_q \rangle) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus g(\langle s_{q-1}, a_{q-1}, u_{q-1}, s_q \rangle, \mathsf{fix}(\tau^M_{0:q-1})) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus g(\langle s_{q-1}, a_{q-1}, u_{q-1} \rangle, \mathsf{fix}(\tau^M_{0:q-1})) \oplus g(\langle s_q \rangle, \mathsf{fix}(\tau^M_{0:q})) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus g(\langle s_{q-1}, a_{q-1}, u_{q-1} \rangle, \mathsf{fix}(\tau^M_{0:q-1})) \oplus g(\langle s_q \rangle, \mathsf{upd}(\mathsf{fix}(\tau^M_{0:q-1}), u_{q-1}, O^{\mathfrak{n}}_{\bullet}(s_{q-1}), O_{\circ}(s_{q-1}), a_{q-1})) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus g(\langle s_{q-1}, a_{q-1}, u_{q-1} \rangle, \mathsf{fix}(\tau^M_{0:q-1})) \oplus g(\langle s_q \rangle, u^{\hookrightarrow}_q) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus \langle \langle s_{q-1}, \mathsf{fix}(\tau^M_{0:q-1}) \rangle, a_{q-1}, \langle s_{q-1}, \mathsf{fix}(\tau^M_{0:q-1}), a_{q-1} \rangle, u_{q-1} \rangle \oplus g(\langle s_q \rangle, u^{\hookrightarrow}_q) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus \langle \langle s_{q-1}, u^{\hookrightarrow}_{0:q-1} \rangle, a_{q-1}, \langle s_{q-1}, u^{\hookrightarrow}_{q-1}, a_{q-1} \rangle, u_{q-1} \rangle \oplus g(\langle s_q \rangle, u^{\hookrightarrow}_q) \\
&= \tau^G_{0:q-1} \oplus \langle a_{q-1}, \langle s_{q-1}, u^{\hookrightarrow}_{q-1}, a_{q-1} \rangle, u_{q-1} \rangle \oplus g(\langle s_q \rangle, u^{\hookrightarrow}_q) \\
&= \tau^G_{0:q-1} \oplus \langle a_{q-1}, \langle s_{q-1}, u^{\hookrightarrow}_{q-1}, a_{q-1} \rangle, u_{q-1}, \langle s_q, u^{\hookrightarrow}_q \rangle \rangle \\
&= \tau^G.
\end{aligned}$$

So if f is surjective for paths of arbitrary length $q-1 \in \mathbb{N}$, f is surjective for paths of length q. Hence, by induction, f is surjective.

f is injective and surjective, hence f is a bijection. □

We write $\simeq$ to indicate equivalence between objects in the RPOMDP and the POSG.

Corollary 1 (Corresponding paths)**.** *f is a bijection between Paths^M and Paths^G, so the set of paths in the RPOMDP is equivalent to the set of paths in the POSG:*

$$\textit{Paths}^M \simeq \textit{Paths}^G,$$

where $\forall \tau^M \in \textit{Paths}^M, \forall \tau^G \in \textit{Paths}^G$.

$$\tau^M \simeq \tau^G \iff f(\tau^M) = \tau^G.$$

We show and prove a bijection between joint histories, as introduced in Appendix A.2. The individual agent and nature histories' bijection proofs follow the same line of reasoning, omitting elements private to the other player.

Proposition 4 (Bijection between joint histories)**.** *Let M be an RPOMDP, and G the POSG of M. There exists a bijection $f^h \colon H^M \to H^G$.*

Proof. Let $H^{M,\ltimes}$ be the joint histories segments for the RPOMDP and let $H^{G,\ltimes}$ be the joint histories segments for the parameterized POSG. Again, we have that $H^M \subseteq H^{M,\ltimes}$ and $H^G \subseteq H^{G,\ltimes}$. Let $h^M = \langle z^{\mathfrak{a}}_{\bullet,0}, z^{\mathfrak{n}}_{\bullet,0}, z_{\circ,0}, a_0, u_0, z^{\mathfrak{a}}_{\bullet,1}, z^{\mathfrak{n}}_{\bullet,1}, z_{\circ,1}, \ldots, z^{\mathfrak{a}}_{\bullet,n}, z^{\mathfrak{n}}_{\bullet,n}, z_{\circ,n} \rangle \in H^M$ and $t \leq n$, then $h^M(t)$ indicates the t-th segment $\langle z^{\mathfrak{a}}_{\bullet,t}, z^{\mathfrak{n}}_{\bullet,t}, z_{\circ,t}, a_t, u_t \rangle$ of h^M. Note that if $t = n$, the segment will only consist of the final observations $\langle z^{\mathfrak{a}}_{\bullet,n}, z^{\mathfrak{n}}_{\bullet,n}, z_{\circ,n} \rangle$. Similarly, let $h^G = \langle \langle z^{\mathfrak{a}}_{\bullet,0}, z_{\circ,0} \rangle, \langle z^{\mathfrak{n}}_{\bullet,0}, z_{\circ,0}, \bot \rangle, a_0, \langle z^{\mathfrak{a}}_{\bullet,0}, z_{\circ,0} \rangle, \langle z^{\mathfrak{n}}_{\bullet,0}, z_{\circ,0}, a_0 \rangle, u_0, \langle z^{\mathfrak{a}}_{\bullet,1}, z_{\circ,1} \rangle, \langle z^{\mathfrak{n}}_{\bullet,1}, z_{\circ,1}, \bot \rangle, \ldots, \langle z^{\mathfrak{a}}_{\bullet,n}, z_{\circ,n} \rangle, \langle z^{\mathfrak{n}}_{\bullet,n}, z_{\circ,n}, \bot \rangle \rangle \in H^G$ and $t \leq n$, then $h^G(t)$ indicates the t-th segment $\langle \langle z^{\mathfrak{a}}_{\bullet,t}, z_{\circ,t} \rangle, \langle z^{\mathfrak{n}}_{\bullet,t}, z_{\circ,t}, \bot \rangle, a_t, \langle z^{\mathfrak{a}}_{\bullet,t}, z_{\circ,t} \rangle, \langle z^{\mathfrak{n}}_{\bullet,t}, z_{\circ,t}, a_t \rangle, u_t \rangle$ of h^G. Note that if $t = n$, the segment will only consist of the final observations $\langle \langle z^{\mathfrak{a}}_{\bullet,n}, z_{\circ,n} \rangle, \langle z^{\mathfrak{n}}_{\bullet,n}, z_{\circ,n}, \bot \rangle \rangle$.

Let $g^h \colon H^{M,\ltimes} \to H^{G,\ltimes}$ defined by:

$$\begin{aligned} g^h(\langle z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}\rangle) &= \langle\langle z^{\mathfrak{a}}_{\bullet}, z_{\circ}\rangle, \langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, \bot\rangle\rangle. \\ g^h(\langle z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a, u\rangle) &= \langle\langle z^{\mathfrak{a}}_{\bullet}, z_{\circ}\rangle, \langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, \bot\rangle, a, \langle z^{\mathfrak{a}}_{\bullet}, z_{\circ}\rangle, \langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a\rangle, u\rangle. \\ g^h(\langle z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a, u\rangle \oplus h') &= g^h(\langle z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a, u\rangle) \oplus g^h(h'). \end{aligned}$$

Let $f^h \colon H^M \to H^G$ defined by:

$$\begin{aligned} f^h(\langle z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}\rangle) &= \langle\langle z^{\mathfrak{a}}_{\bullet}, z_{\circ}\rangle, \langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, \bot\rangle\rangle. \\ f^h(\langle z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a, u\rangle) &= \langle\langle z^{\mathfrak{a}}_{\bullet}, z_{\circ}\rangle, \langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, \bot\rangle, a, \langle z^{\mathfrak{a}}_{\bullet}, z_{\circ}\rangle, \langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a\rangle, u\rangle. \\ f^h(\langle z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a, u\rangle \oplus h') &= f^h(\langle z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a, u\rangle) \oplus g^h(h'). \end{aligned}$$

Note that the results of f^h and g^h are in H^G and $H^{G,\ltimes}$ by construction. Also, note that these function definitions are similar to those for paths.

We show that f^h is a bijection. We first show f^h is injective, so we show that:

$$\forall h^{1,M}, h^{2,M} \in H^M . h^{1,M} \neq h^{2,M} \implies f^h(h^{1,M}) \neq f^h(h^{2,M}).$$

Given arbitrary $h^{1,M}, h^{2,M} \in H^M$, we distinguish between the histories with a superscript $1, 2$ respectively. Assume $h^{1,M} \neq h^{2,M}$. If $h^{1,M}$ and $h^{2,M}$ do not have the same horizon length, then neither do $f^h(h^{1,M})$ and $f^h(h^{2,M})$. Then trivially, $f^h(h^{1,M}) \neq f^h(h^{2,M})$.

Assume $f^h(h^{1,M})$ and $f^h(h^{2,M})$ have the same horizon length n. Then $\exists t \leq n$ where $f^h(h^{1,M})$ and $f^h(h^{2,M})$ deviate, so $h^{1,M}(t) \neq h^{2,M}(t)$. Let q be the smallest number where the histories deviate. So $\forall t < q . h^{1,M}(t) = h^{2,M}(t)$ and $h^{1,M}(q) \neq h^{2,M}(q)$. Assume $q < n$. Then we know $\langle z^{\mathfrak{a},1}_{\bullet,q}, z^{\mathfrak{n},1}_{\bullet,q}, z^1_{\circ,q}, a^1_q, u^1_q\rangle \neq \langle z^{\mathfrak{a},2}_{\bullet,q}, z^{\mathfrak{n},2}_{\bullet,q}, z^2_{\circ,q}, a^2_q, u^2_q\rangle$, which comes down to: $z^{\mathfrak{a},1}_{\bullet,q} \neq z^{\mathfrak{a},2}_{\bullet,q} \vee z^{\mathfrak{n},1}_{\bullet,q} \neq z^{\mathfrak{n},2}_{\bullet,q} \vee z^1_{\circ,q} \neq z^2_{\circ,q} \vee a^1_q \neq a^2_q \vee u^1_q \neq u^2_q$.

$$\begin{aligned} f^h(h^{1,M}) &= f^h(h^{1,M}(1) \oplus h_1^{M\prime}) \\ &= h_1^G(1) \oplus g^h(h^{1,M\prime}). \end{aligned}$$

Unfold g^h until q:

$$= \bigoplus_{t=0}^{q-1} (h_1^G(t)) \oplus \langle\langle z^{\mathfrak{a},1}_{\bullet,q}, z^1_{\circ,q}\rangle, \langle z^{\mathfrak{n},1}_{\bullet,q}, z^1_{\circ,q}, \bot\rangle, a^1_q, \langle z^{\mathfrak{a},1}_{\bullet,q}, z^1_{\circ,q}\rangle, \langle z^{\mathfrak{n},1}_{\bullet,q}, z_{\circ,q}, a^1_q\rangle, u^1_q\rangle \oplus g^h(h^{1,M\prime\prime}).$$

Since $z^{\mathfrak{a},1}_{\bullet,q} \neq z^{\mathfrak{a},2}_{\bullet,q} \vee z^{\mathfrak{n},1}_{\bullet,q} \neq z^{\mathfrak{n},2}_{\bullet,q} \vee z^1_{\circ,q} \neq z^2_{\circ,q} \vee a^1_q \neq a^2_q \vee u^1_q \neq u^2_q$:

$$\begin{aligned} &\neq \bigoplus_{t=0}^{q-1} (h_1^G(t)) \oplus \langle\langle z^{\mathfrak{a},2}_{\bullet,q}, z^2_{\circ,q}\rangle, \langle z^{\mathfrak{n},2}_{\bullet,q}, z^2_{\circ,q}, \bot\rangle, a^2_q, \langle z^{\mathfrak{a},2}_{\bullet,q}, z^2_{\circ,q}\rangle, \langle z^{\mathfrak{n},2}_{\bullet,q}, z_{\circ,q}, a^2_q\rangle, u^2_q\rangle \oplus g^h(h^{2,M\prime\prime}) \\ &= \bigoplus_{t=0}^{q-1} (h^{2,G}(t)) \oplus \langle\langle z^{\mathfrak{a},2}_{\bullet,q}, z^2_{\circ,q}\rangle, \langle z^{\mathfrak{n},2}_{\bullet,q}, z^2_{\circ,q}, \bot\rangle, a^2_q, \langle z^{\mathfrak{a},2}_{\bullet,q}, z^2_{\circ,q}\rangle, \langle z^{\mathfrak{n},2}_{\bullet,q}, z_{\circ,q}, a^2_q\rangle, u^2_q\rangle \oplus g^h(h^{2,M\prime\prime}). \end{aligned}$$

Fold g^h until 1:

$$\begin{aligned} &= h^{2,G}(1) \oplus g^h(h^{2,M\prime}) \\ &= f^h(h^{2,M}(1) \oplus h_2^{M\prime}) \\ &= f^h(h^{2,M}). \end{aligned}$$

If $q = n$, then the same result follows by removing everything after $\langle z^{\mathfrak{n},1}_{\bullet,q}, z^1_{\circ,q}, \bot\rangle$, and $\langle z^{\mathfrak{n},2}_{\bullet,q}, z^2_{\circ,q}, \bot\rangle$. We thus have that $f^h(h^{1,M}) \neq f^h(h^{2,M})$, so f^h is injective.

Next, we show that f^h is surjective, so we show that:

$$\forall h^G \in H^G, \exists h^M \in H^M . f^h(h^M) = h^G.$$

Take arbitrary $h^G \in H^G$. By construction of H^G, O^G (see Appendix A.2) is surjective, so we know $\exists \tau^G \in \text{Paths}^G, O^G(\tau^G) = h^G$. Take $\tau^G \in \text{Paths}^G$ such that $O^G(\tau^G) = h^G$. Let $\tau^M \in \text{Paths}^M$ be the corresponding path in the RPOMDP. So $f(\tau^M) = \tau^G$. Then $h^M = O^M(\tau^M) \in H^M$.

We proof $f^h(h^M) = h^G$ by contradiction. Assume $f^h(h^M) = h_2^G \neq h^G$. By construction of f, we know $|\tau^M| = |\tau^G|$. Then, by construction of O^M, O^G, and f^h, which each map a segment to a segment, we know $|h_2^G| = |h^G|$.

Let n be the horizon length of h_2^G and h^G. Then $\exists t \in \mathbb{N}$ where h_2^G and h^G deviate, so $h^G(t) \neq h_2^G(t)$. Let q be such a number where the horizons deviate. So $h^G(q) \neq h_2^G(q)$. Note that $q \geq 1$, since there is only one initial state, therefore $h^G(0) = h_2^G(0)$. Assume $q < n$ and let $\tau^G(q) = \langle\langle s_q, u_q^{\hookrightarrow}\rangle, a_q, \langle s_q, u_q^{\hookrightarrow}, a_q\rangle, u_q\rangle$. Then by the definition and bijectivity of f, we know $\tau^M(q) = \langle s_q, a_q, u_q\rangle$. Furthermore, by construction, we know that $f^h, g^h, O^G, O^{G,\ltimes}, O^M$, and $O^{M,\ltimes}$ all apply on the segments of the paths or histories separately.

$$
\begin{aligned}
h^G(q) &= O^G(\tau^G(q)) \\
&= O^G(\langle\langle s_q, u_q^{\hookrightarrow}\rangle, a_q, \langle s_q, u_q^{\hookrightarrow}, a_q\rangle, u_q\rangle) \\
&= \langle\langle O_\bullet^{\mathfrak{a}}(s_q), O_\circ(s_q)\rangle, \langle O_\bullet^{\mathfrak{n}}(s_q), O_\circ(s_q), \bot\rangle, a_q, \langle O_\bullet^{\mathfrak{a}}(s_q), O_\circ(s_q)\rangle, \langle O_\bullet^{\mathfrak{n}}(s_q), O_\circ(s_q), a_q\rangle, u_q\rangle. \\
h_2^G(q) &= f^h(O^M(\tau^M(q))) \\
&= f^h(O^M(\langle s_q, a_q, u_q\rangle)) \\
&= f^h(\langle O_\bullet^{\mathfrak{a}}(s_q), O_\bullet^{\mathfrak{n}}(s_q), O_\circ(s_q), a_q, u_q\rangle) \\
&= \langle\langle O_\bullet^{\mathfrak{a}}(s_q), O_\circ(s_q)\rangle, \langle O_\bullet^{\mathfrak{n}}(s_q), O_\circ(s_q), \bot\rangle, a_q, \langle O_\bullet^{\mathfrak{a}}(s_q), O_\circ(s_q)\rangle, \langle O_\bullet^{\mathfrak{n}}(s_q), O_\circ(s_q), a_q\rangle, u_q\rangle \\
&= h^G(q).
\end{aligned}
$$

Hence $\exists t \in \mathbb{N} : h^G(t) \neq h_2^G(t)$ is false. If $q = n$, the same result follows by removing everything from a_q. We hence get that $f^h(h^M) = h^G$, so $\exists h^M \in H^M, f^h(h^M) = h^G$, therefore f^h is surjective.

f^h is injective and surjective, hence f^h is a bijection. □

Lemma 3 (Bijection between agent and nature histories)**.** *Following Proposition 4, we get bijections for the agent and nature histories by omitting the private objects of the other player. Let $g^{\mathfrak{a},h} \colon H^{\mathfrak{a},M,\ltimes} \to H^{\mathfrak{a},G,\ltimes}$ defined by:*

$$
\begin{aligned}
g^{\mathfrak{a},h}(\langle z_\bullet^{\mathfrak{a}}, z_\circ\rangle) &= \langle\langle z_\bullet^{\mathfrak{a}}, z_\circ\rangle\rangle. \\
g^{\mathfrak{a},h}(\langle z_\bullet^{\mathfrak{a}}, z_\circ, a\rangle) &= \langle\langle z_\bullet^{\mathfrak{a}}, z_\circ\rangle, a, \langle z_\bullet^{\mathfrak{a}}, z_\circ\rangle\rangle. \\
g^{\mathfrak{a},h}(\langle z_\bullet^{\mathfrak{a}}, z_\circ, a, h'\rangle) &= g^{\mathfrak{a},h}(\langle z_\bullet^{\mathfrak{a}}, z_\circ, a\rangle) \oplus g^{\mathfrak{a},h}(h').
\end{aligned}
$$

Let $f^{\mathfrak{a},h} \colon H^{\mathfrak{a},M} \to H^{\mathfrak{a},G}$ defined by:

$$
\begin{aligned}
f^{\mathfrak{a},h}(\langle z_\bullet^{\mathfrak{a}}, z_\circ\rangle) &= \langle\langle z_\bullet^{\mathfrak{a}}, z_\circ\rangle\rangle. \\
f^{\mathfrak{a},h}(\langle z_\bullet^{\mathfrak{a}}, z_\circ, a\rangle) &= \langle\langle z_\bullet^{\mathfrak{a}}, z_\circ\rangle, a, \langle z_\bullet^{\mathfrak{a}}, z_\circ\rangle\rangle. \\
f^{\mathfrak{a},h}(\langle z_\bullet^{\mathfrak{a}}, z_\circ, a, h'\rangle) &= f^{\mathfrak{a},h}(\langle z_\bullet^{\mathfrak{a}}, z_\circ, a\rangle) \oplus g^{\mathfrak{a},h}(h').
\end{aligned}
$$

$f^{\mathfrak{a},h}$ is a bijection.

Let $g^{\mathfrak{n},h} \colon H^{\mathfrak{n},M,\ltimes} \to H^{\mathfrak{n},G,\ltimes}$ defined by:

$$
\begin{aligned}
g^{\mathfrak{n},h}(\langle z_\bullet^{\mathfrak{n}}, z_\circ\rangle) &= \langle\langle z_\bullet^{\mathfrak{n}}, z_\circ, \bot\rangle\rangle. \\
g^{\mathfrak{n},h}(\langle z_\bullet^{\mathfrak{n}}, z_\circ, a, u\rangle) &= \langle\langle z_\bullet^{\mathfrak{n}}, z_\circ, \bot\rangle, \langle z_\bullet^{\mathfrak{n}}, z_\circ, a\rangle, u\rangle. \\
g^{\mathfrak{n},h}(\langle z_\bullet^{\mathfrak{n}}, z_\circ, a, u, h'\rangle) &= g^{\mathfrak{n},h}(\langle z_\bullet^{\mathfrak{n}}, z_\circ, a, u\rangle) \oplus g^{\mathfrak{n},h}(h').
\end{aligned}
$$

Let $f^{\mathfrak{n},h} \colon H^{\mathfrak{n},M} \to H^{\mathfrak{n},G}$ defined by:

$$
\begin{aligned}
f^{\mathfrak{n},h}(\langle z_\bullet^{\mathfrak{n}}, z_\circ\rangle) &= \langle\langle z_\bullet^{\mathfrak{n}}, z_\circ, \bot\rangle\rangle. \\
f^{\mathfrak{n},h}(\langle z_\bullet^{\mathfrak{n}}, z_\circ, a, u\rangle) &= \langle\langle z_\bullet^{\mathfrak{n}}, z_\circ, \bot\rangle, \langle z_\bullet^{\mathfrak{n}}, z_\circ, a\rangle, u\rangle. \\
f^{\mathfrak{n},h}(\langle z_\bullet^{\mathfrak{n}}, z_\circ, a, u, h'\rangle) &= f^{\mathfrak{n},h}(\langle z_\bullet^{\mathfrak{n}}, z_\circ, a, u\rangle) \oplus g^{\mathfrak{n},h}(h').
\end{aligned}
$$

$f^{\mathfrak{n},h}$ is a bijection.

Corollary 2 (Corresponding histories)**.** *f^h is a bijection between H^M and H^G, so the set of histories in the RPOMDP is equivalent to the set of histories in the parameterized POSG:*

$$H^M \simeq H^G,$$

where $\forall h^M \in H^M, \forall h^G \in H^G$.

$$h^M \simeq h^G \iff f^h(h^M) = h^G.$$

Similarly:

$$H^{\mathfrak{a},M} \simeq H^{\mathfrak{a},G},$$

where $\forall h^{\mathfrak{a},M} \in H^{\mathfrak{a},M}, \forall h^{\mathfrak{a},G} \in H^{\mathfrak{a},G}$.

$$h^{\mathfrak{a},M} \simeq h^{\mathfrak{a},G} \iff f^{\mathfrak{a},h}(h^{\mathfrak{a},M}) = h^{\mathfrak{a},G}.$$

And:

$$H^{\mathfrak{n},M} \simeq H^{\mathfrak{n},G},$$

where $\forall h^{\mathfrak{a},M} \in H^{\mathfrak{a},M}, \forall h^{\mathfrak{a},G} \in H^{\mathfrak{a},G}$.

$$h^{\mathfrak{a},M} \simeq h^{\mathfrak{a},G} \iff f^{\mathfrak{a},h}(h^{\mathfrak{a},M}) = h^{\mathfrak{a},G}.$$

Proposition 2 from the main text is a direct corollary of the bijections between histories established above. For completeness, we repeat the proposition here.

Proposition 2 (Bijection between policies)**.** *Let $f^\pi \colon \Pi^M \to \Pi^G$ defined by:*

$$f^\pi(\pi^M)(h^{\mathfrak{a},G}) = \pi^M((f^{\mathfrak{a},h})^{-1}(h^{\mathfrak{a},G})),$$

then f^π is a bijection.

Let $f^\theta \colon \Theta^M \to \Theta^G$ defined by:

$$f^\theta(\theta^M)(h^{\mathfrak{n},G}, \langle z^{\mathfrak{n}}_\bullet, z_\circ, a\rangle) = \theta^M((f^{\mathfrak{n},h})^{-1}(h^{\mathfrak{n},G}), a),$$

then f^θ is a bijection.

Corollary 3 (Corresponding policies)**.** *f^π is a bijection between agent policies, so the set of agent policies in the RPOMDP is equivalent to the set of agent policies in the parameterized POSG:*

$$\Pi^M \simeq \Pi^G,$$

where $\forall \pi^M \in \Pi^M, \forall \pi^G \in \Pi^G$.

$$\pi^M \simeq \pi^G \iff f^\pi(\pi^M) = \pi^G.$$

Similarly:

$$\Theta^M \simeq \Theta^G,$$

where $\forall \theta^M \in \Theta^M, \forall \theta^G \in \Theta^G$.

$$\theta^M \simeq \theta^G \iff f^\theta(\theta^M) = \theta^G.$$

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:*

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

Proof. We prove $R(\tau^M) = \mathcal{R}(\tau^G)$ for corresponding paths τ^M, τ^G. By definition of corresponding paths, we know that $\forall t \in \mathbb{N}. \tau^G(t) = g(\tau^M(t), u^{\hookrightarrow}_{t-1})$.

$$\begin{aligned} R(\tau^M) &= \sum_{t \in \mathbb{N}} R(\tau^M(t)) \\ &= \sum_{t \in \mathbb{N}} R(s_t, a_t, u_t) \\ &= \sum_{t \in \mathbb{N}} R(s_t, a_t) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{t \in \mathbb{N}} \mathcal{R}(\langle s_t, u_{t-1}^{\hookrightarrow} \rangle, a_t) \\
&= \sum_{t \in \mathbb{N}} \mathcal{R}(\langle s_t, u_{t-1}^{\hookrightarrow} \rangle, a_t, \langle s_t, u_{t-1}^{\hookrightarrow}, a_t \rangle, u_t) \\
&= \sum_{t \in \mathbb{N}} \mathcal{R}(\tau^G(t)) \\
&= \mathcal{R}(\tau^G).
\end{aligned}
$$

Now, since joint corresponding policies π^M, θ^M and π^G, θ^G lead to the same distribution over corresponding paths, we know that $V_\phi^{\pi^M, \theta^M} = V_\phi^{\pi^G, \theta^G}$. □

F.1 Mixed policies

Given the bijection between histories and stochastic policies, we can define a bijection between mixed policies similar to the one between stochastic policies. Note that we apply the bijection between stochastic policies to deterministic policies.

Proposition 5 (Bijection between mixed policies)**.** *Let $f^{\pi,mix} \colon \Pi^{M,mix} \to \Pi^{G,mix}$ defined by:*

$$f^{\pi,mix}(\pi^{M,mix})(\pi^{G,det}) = \pi^{M,mix}((f^{\pi})^{-1}(\pi^{G,det})),$$

then $f^{\pi,mix}$ is a bijection.
Let $f^{\theta,mix} \colon \Theta^{M,mix} \to \Theta^{G,mix}$ defined by:

$$f^{\theta,mix}(\theta^{M,mix})(\theta^{G,det}) = \theta^{M,mix}((f^{\theta})^{-1}(\theta^{G,det})),$$

then $f^{\theta,mix}$ is a bijection.

Corollary 4 (Corresponding mixed policies)**.** *$f^{\pi,mix}$ is a bijection between agent policies, so the set of mixed agent policies in the RPOMDP is equivalent to the set of mixed agent policies in the parameterized POSG:*

$$\Pi^{M,mix} \simeq \Pi^{G,mix},$$

where $\forall \pi^{M,mix} \in \Pi^{M,mix}, \forall \pi^{G,mix} \in \Pi^{G,mix}$.

$$\pi^{M,mix} \simeq \pi^{G,mix} \iff f^{\pi,mix}(\pi^{M,mix}) = \pi^{G,mix}.$$

Similarly:

$$\Theta^{M,mix} \simeq \Theta^{G,mix},$$

where $\forall \theta^{M,mix} \in \Theta^{M,mix}, \forall \theta^{G,mix} \in \Theta^{G,mix}$.

$$\theta^{M,mix} \simeq \theta^{G,mix} \iff f^{\theta,mix}(\theta^{M,mix}) = \theta^{G,mix}.$$

Theorem 4 (Equivalent values mixed policies)**.** *Let M be an RPOMDP, and G the POSG of M. Let $\pi^{M,mix} \in \Pi^{M,mix}, \pi^{G,mix} = f^{\pi,mix}(\pi^{M,mix}) \in \Pi^{G,mix}$ be corresponding agent policies, and $\theta^{M,mix} \in \Theta^{M,mix}, \theta^{G,mix} = f^{\theta,mix}(\theta^{M,mix}) \in \Theta^{G,mix}$ be corresponding nature policies. Then, their values for the RPOMDP and POSG coincide:*

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

The proof follows the same steps as for Theorem 2 for stochastic policies, where the distribution over corresponding paths now follows from the same distribution over corresponding deterministic policies, which in turn leads to the same distributions over corresponding paths.

F.2 Nature first

As explained in Appendix E, when reasoning with nature first semantics, the paths of the POSGs change. The bijections in this appendix start from the bijection between paths. Below, we give the bijections for the nature first semantics that differ from the bijections for the agent first semantics. The bijection proofs follow the same steps in the nature first case as in the agent first case. We show the adjusted proof for the path bijection to illustrate how to deal with the delayed observation of the last agent action.

Lemma 4 (Nature first bijection between paths)**.** *Let M be an RPOMDP, and G the POSG of M. There exists a bijection $f \colon Paths^M \to Paths^G$*

Proof. Let $\text{Paths}^{M,\ltimes} \subseteq (S \times A \times \boldsymbol{U})^* \times S^?$ with $? \in \{0,1\}$ be the set of all path segments in the RPOMDP, and let $\text{Paths}^{G,\ltimes} \subseteq \mathcal{S}^{\mathfrak{n}} \times \mathcal{A}^{\mathfrak{n}} \times \mathcal{S}^{\mathfrak{a}} \times \mathcal{A}^{\mathfrak{a}})^* \times (\mathcal{S}^{\mathfrak{n}})^?$ with $? \in \{0,1\}$ be the set of all path segments in the POSG. With path segment we mean that the path can starts at any time steps $t \in \mathbb{N}$ and can end at any time step $t' \in \mathbb{N}, t <= t'$. The optional last state is only used for path segments until the horizon. Note that $\text{Paths}^M \subseteq \text{Paths}^{M,\ltimes}$ and $\text{Paths}^G \subseteq \text{Paths}^{G,\ltimes}$. Let $\tau^M = \langle s_0, a_0, u_0, s_1, \ldots, s_n \rangle \in \text{Paths}^M$ and $t \leq n$, then $\tau^M(t)$ indicates the t-th segment $\langle s_t, a_t, u_t \rangle$ of τ^M. Note that if $t = n$, the segment will only consist of the final state $\langle s_n \rangle$. Similarly, let $\tau^G = \langle s_0^{\mathfrak{n}}, a_0^{\mathfrak{n}}, s_0^{\mathfrak{a}}, a_0^{\mathfrak{a}}, s_1^{\mathfrak{n}}, \ldots, s_n^{\mathfrak{n}} \rangle = \langle \langle s_0, u^{\hookrightarrow}, \bot \rangle, u_0, \langle s_0, u^{\bot}, u_0 \rangle, a_0, \langle s_1, u_1^{\hookrightarrow}, a_0 \rangle, \ldots, \langle s_n, u_n^{\hookrightarrow}, a_{n-1} \rangle \rangle \in \text{Paths}^G$ and $t \leq n$, then $\tau^G(t)$ indicates the t-th segment $\langle s_t^{\mathfrak{n}}, a_t^{\mathfrak{n}}, s_t^{\mathfrak{a}}, a_t^{\mathfrak{a}} \rangle = \langle \langle s_t, u_t^{\hookrightarrow}, a_{t-1} \rangle, u_t, \langle s_t, u_t^{\hookrightarrow}, u_t \rangle, a_t \rangle$ of τ^G. Note that if $t = n$, the segment will only consist of the final nature state $\langle s_n^{\mathfrak{a}} \rangle = \langle \langle s_n, u_n^{\hookrightarrow}, a_{n-1} \rangle \rangle$.

Let $g \colon \text{Paths}^{M,\ltimes} \times \boldsymbol{U}^{\hookrightarrow} \times A \hookrightarrow \text{Paths}^{G,\ltimes}$ defined by:

$$
\begin{aligned}
g(\langle s \rangle, u^{\hookrightarrow}, a) &= \langle \langle s, u^{\hookrightarrow}, a' \rangle \rangle. \\
g(\langle s, a, u \rangle, u^{\hookrightarrow}, a') &= \begin{cases} \langle \langle s, u^{\hookrightarrow}, a' \rangle, u, \langle s, u^{\hookrightarrow}, u \rangle, a \rangle & \text{if } u \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookrightarrow}), \\ \bot & \text{otherwise.} \end{cases} \\
g(\langle s, a, u \rangle \oplus \tau^{M'}, u^{\hookrightarrow}, a') &= \begin{cases} g(\langle s, a, u \rangle, u^{\hookrightarrow}, a') \oplus g(\tau^{M'}, \mathsf{upd}(u^{\hookrightarrow}, u, O_{\bullet}^{\mathfrak{n}}(s), O_{\circ}(s), a), a) & \text{if } u \in \boldsymbol{U}^{\mathcal{P}}(u^{\hookrightarrow}), \\ \bot & \text{otherwise.} \end{cases}
\end{aligned}
$$

Let $f \colon \text{Paths}^M \to \text{Paths}^G$ defined by:

$$
\begin{aligned}
f(\langle s \rangle) &= \langle \langle s, u^{\bot}, \bot \rangle \rangle. \\
f(\langle s, a, u \rangle) &= \langle \langle s, u^{\bot}, \bot \rangle, u, \langle s, u^{\bot}, u \rangle, a \rangle. \\
f(\langle s, a, u \rangle \oplus \tau^{M'}) &= f(\langle s, a, u \rangle) \oplus g(\tau^{M'}, \mathsf{upd}(u^{\bot}, u, O_{\bullet}^{\mathfrak{n}}(s), O_{\circ}(s), a), a).
\end{aligned}
$$

Where $u^{\bot} \in \boldsymbol{U}^{\hookrightarrow}$ is the totally undefined function. Note that the results of f and g are in Paths^G and $\text{Paths}^{G,\ltimes}$ by construction. Also, note that any call to g that originated from a call in f will have a result by construction.

We show that f is a bijection, meaning f is injective and surjective. We first show f is injective, so we show that:

$$
\forall \tau^{1,M}, \tau^{2,M} \in \text{Paths}^M . \tau^{1,M} \neq \tau^{2,M} \implies f(\tau^{1,M}) \neq f(\tau^{2,M}).
$$

Given arbitrary $\tau^{1,M}, \tau^{2,M} \in \text{Paths}^M$, we distinguish between the paths with superscripts 1 and 2, respectively. Assume $\tau^{1,M} \neq \tau^{2,M}$. If $\tau^{1,M}$ and $\tau^{2,M}$ do not have the same horizon length, then neither do $f(\tau^{1,M})$ and $f(\tau^{2,M})$. Then trivially, $f(\tau^{1,M}) \neq f(\tau^{2,M})$.

Assume $\tau^{1,M}$ and $\tau^{2,M}$ have the same horizon length n. Then $\exists t \leq n$ where $\tau^{1,M}$ and $\tau^{2,M}$ deviate, so $\tau^{1,M}(t) \neq \tau^{2,M}(t)$. Let q be the smallest number where the paths deviate. So $\forall t < q . \tau^{1,M}(t) = \tau^{2,M}(t)$ and $\tau^{1,M}(q) \neq \tau^{2,M}(q)$. Assume $q < n$. Then we know $\langle s_q^1, a_q^1, u_q^1 \rangle \neq \langle s_q^2, a_q^2, u_q^2 \rangle$, which comes down to: $s_q^1 \neq s_q^2 \vee a_q^1 \neq a_q^2 \vee u_q^1 \neq u_q^2$.

$$
\begin{aligned}
f(\tau^{1,M}) &= f(\tau^{1,M}(1) \oplus \tau^{1,M'}) \\
&= \tau^{1,G}(1) \oplus g(\tau^{1,M'}, \mathsf{upd}(u^{\bot}, u_0^1, O_{\bullet}^{\mathfrak{n}}(s_0^1), O_{\circ}(s_0^1), a_0^1), a_0^1).
\end{aligned}
$$

Unfold g until q:

$$
= \bigoplus_{t=0}^{q} (\tau^{1,G}(t)) \oplus \langle \langle s_q^1, \mathsf{fix}(\tau_{0:q}^{1,M}), a_{q-1}^1 \rangle, u_q^1, \langle s_q^1, \mathsf{fix}(\tau_{0:q}^{1,M}), u_q^1 \rangle, a_q^1 \rangle \oplus g(\tau^{1,M''}, \mathsf{fix}(\tau_{0:q+1}^{1,M}), a_q^1)
$$

Since $s_q^1 \neq s_q^2 \vee a_q^1 \neq a_q^2 \vee u_q^1 \neq u_q^2$:

$$
\begin{aligned}
&\neq \bigoplus_{t=0}^{q} (\tau^{1,G}(t)) \oplus \langle \langle s_q^2, \mathsf{fix}(\tau_{0:q}^{1,M}), a_{q-1}^1 \rangle, u_q^2, \langle s_q^2, \mathsf{fix}(\tau_{0:q}^{1,M}), u_q^2 \rangle, a_q^2 \rangle \oplus g(\tau^{2,M''}, \mathsf{fix}(\tau_{0:q+1}^{1,M}), a_q^2) \\
&= \bigoplus_{t=0}^{q} (\tau^{2,G}(t)) \oplus \langle \langle s_q^2, \mathsf{fix}(\tau_{0:q}^{2,M}), a_{q-1}^2 \rangle, u_q^2, \langle s_q^2, \mathsf{fix}(\tau_{0:q}^{2,M}), u_q^2 \rangle, a_q^2 \rangle \oplus g(\tau^{2,M''}, \mathsf{fix}(\tau_{0:q+1}^{2,M}), a_q^2).
\end{aligned}
$$

Fold g until 1:

$$\begin{aligned}
&= \tau^{2,G}(1) \oplus g(\tau^{2,M\prime}, \mathsf{upd}(u^{\perp}, u_0^2, O_\bullet^{\mathfrak{n}}(s_0^2), O_\circ(s_0^2), a_0^2), a_0^2) \\
&= f(\tau^{2,M}(1) \oplus \tau^{2,M\prime}) \\
&= f(\tau^{2,M}).
\end{aligned}$$

If $q = n$, then the same result follows by removing everything after $\langle s_q^1, \mathsf{fix}(\tau_{0:q}^{1,M}), a_{q-1}^1\rangle, \langle s_q^2, \mathsf{fix}(\tau_{0:q}^{1,M}), a_{q-1}^1\rangle$, and $\langle s_q^2, \mathsf{fix}(\tau_{0:q}^{2,M}), a_{q-1}^2\rangle$. We thus have that $f(\tau^{1,M}) \neq f(\tau^{2,M})$, so f is injective.

Next, we show that f is surjective, so we show that:

$$\forall \tau^G \in \text{Paths}^G, \exists \tau^M \in \text{Paths}^M . f(\tau^M) = \tau^G.$$

We show this holds by induction on the horizon length of the $\tau^G \in \text{Paths}^G$. We write the length of τ^G as $|\tau^G|$.

Assume $|\tau^G| = 0$. Then $\tau^G = \langle s_I, u^{\perp}\rangle$. We have that for $\langle s_I\rangle \in \text{Paths}^M$, $f(\langle s_I\rangle) = \langle\langle s_I, u^{\perp}\rangle\rangle = \tau^G$. So for paths of horizon length 0, f is surjective.

Now assume we know, given $q \in \mathbb{N}, q \geq 1$, that:

$$\forall \tau^G \in \text{Paths}^G . |\tau^G| = q - 1 \implies \exists \tau^M \in \text{Paths}^M : f(\tau^M) = \tau^G.$$

Take arbitrary $\tau^G \in \text{Paths}^G$ with horizon length $|\tau^G| = q$. Then we have $\tau^G = \tau_{0:q-1}^G \oplus \langle u_{q-1}, \langle s_{q-1}, u_{q-1}^{\hookrightarrow}, u_{q-1}\rangle, a_{q-1}, \langle s_q, u_q^{\hookrightarrow}, a_{q-1}\rangle\rangle$. Then $\tau_{0:q-1}^G \in \text{Paths}^G$ and $|\tau_{0:q-1}^G| = q - 1$. By assumption, we get that:

$$\exists \tau_{0:q-1}^M \in \text{Paths}^M, f(\tau_{0:q-1}^M) = \tau_{0:q-1}^G.$$

Let $\tau_{0:q-1}^M \in \text{Paths}^M$ such that $f(\tau_{0:q-1}^M) = \tau_{0:q-1}^G$. We then know that in τ^G:

$$\forall t < q . u_t^{\hookrightarrow} = \mathsf{fix}(\tau_{0:t}^M).$$

And, by Definition 18 and the definition of Paths^G, that:

$$u_q^{\hookrightarrow} = \mathsf{upd}(\mathsf{fix}(\tau_{0:q-1}^M), u_{q-1}, O_\bullet^{\mathfrak{n}}(s_{q-1}), O_\circ(s_{q-1}), a_{q-1}).$$

Let $\langle\langle s_{q-2}, \mathsf{fix}(\tau_{0:q-2}^M), a_{q-3}\rangle, u_{q-2}, \langle s_{q-2}, \mathsf{fix}(\tau_{0:q-2}^M), u_{q-2}\rangle, a_{q-2}, \langle s_{q-1}, \mathsf{fix}(\tau_{0:q-1}^M), a_{q-2}\rangle\rangle$ be the last two segments of $\tau_{0:q-1}^G$. Then by definition and injectivity of f, we know that the last two segments of $\tau_{0:q-1}^M$ are $\langle s_{q-2}, a_{q-2}, u_{q-2}, s_{q-1}\rangle$.

Now, by definition Definition 18 and the definition of Paths^G, we know that:

$$\begin{gathered}
\tau^G = \tau_{0:q-1}^G \oplus \langle u_{q-1}, \langle s_{q-1}, u_{q-1}^{\hookrightarrow}, u_{q-1}\rangle, a_{q-1}, \langle s_q, u_q^{\hookrightarrow}, a_{q-1}\rangle\rangle \in \text{Paths}^G \\
\iff \\
\tau_{0:q-1}^G \in \text{Paths}^G \wedge \mathcal{T}^{\mathfrak{n}}(\langle s_{q-1}, u_{q-1}^{\hookrightarrow}, a_{q-2}\rangle, u_{q-1}, \langle s_{q-1}, u_{q-1}^{\hookrightarrow}, u_{q-1}\rangle) > 0 \\
\wedge \mathcal{T}^{\mathfrak{a}}(\langle s_{q-1}, u_{q-1}^{\hookrightarrow}), u_{q-1}\rangle, a_{q-1}, \langle s_q, u_q^{\hookrightarrow}, a_{q-1}\rangle) > 0 \\
\iff \\
\tau_{0:q-1}^G \in \text{Paths}^G \wedge u_{q-1} \in \boldsymbol{U}^{\mathcal{P}}(u_{q-1}^{\hookrightarrow}) \wedge \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) > 0 \\
\iff \\
\tau_{0:q-1}^G \in \text{Paths}^G \wedge u_{q-1} \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(\tau_{0:q-1}^M)) \wedge \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) > 0.
\end{gathered}$$

So, since $\tau^G \in \text{Paths}^G$, we know $u_{q-1} \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(\tau_{0:q-1}^M))$ and $\boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) > 0$, which are the restrictions for $\tau^M = \tau_{0:q-1}^M \oplus \langle a_{q-1}, u_{q-1}, s_q\rangle \in \text{Paths}^M$ to hold.

$$\begin{aligned}
f(\tau^M) &= f(\tau_{0:q-1}^M \oplus \langle a_{q-1}, u_{q-1}, s_q\rangle) \\
&= \bigoplus_{t=0}^{q-2} \tau_{0:q-1}^G(t) \oplus g(\langle s_{q-1}, a_{q-1}, u_{q-1}, s_q\rangle, \mathsf{fix}(\tau_{0:q-1}^M), a_{q-2})
\end{aligned}$$

$$
\begin{aligned}
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus g(\langle s_{q-1}, a_{q-1}, u_{q-1}\rangle, \mathsf{fix}(\tau^M_{0:q-1}), a_{q-2}) \oplus g(\langle s_q\rangle, \mathsf{fix}(\tau^M_{0:q}), a_{q-1}) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus g(\langle s_{q-1}, a_{q-1}, u_{q-1}\rangle, \mathsf{fix}(\tau^M_{0:q-1}), a_{q-2}) \oplus g(\langle s_q\rangle, \mathsf{upd}(\mathsf{fix}(\tau^M_{0:q-1}), u_{q-1}, O^{\mathfrak{n}}_\bullet(s_{q-1}), O_\circ(s_{q-1}), a_{q-1}), a_{q-1}) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus g(\langle s_{q-1}, a_{q-1}, u_{q-1}\rangle, \mathsf{fix}(\tau^M_{0:q-1}), a_{q-2}) \oplus g(\langle s_q\rangle, u^{\hookrightarrow}_q, a_{q-1}) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus \langle\langle s_{q-1}, \mathsf{fix}(\tau^M_{0:q-1}), a_{q-2}\rangle, u_{q-1}, \langle s_{q-1}, \mathsf{fix}(\tau^M_{0:q-1}), u_{q-1}\rangle, a_{q-1}\rangle \oplus g(\langle s_q\rangle, u^{\hookrightarrow}_q, a_{q-1}) \\
&= \bigoplus_{t=0}^{q-2} \tau^G_{0:q-1}(t) \oplus \langle\langle s_{q-1}, u^{\hookrightarrow}_{q-1}, a_{q-2}\rangle, u_{q-1}, \langle s_{q-1}, u^{\hookrightarrow}_{q-1}, u_{q-1}\rangle, a_{q-1}\rangle \oplus g(\langle s_q\rangle, u^{\hookrightarrow}_q, a_{q-1}) \\
&= \tau^G_{0:q-1} \oplus \langle u_{q-1}, \langle s_{q-1}, u^{\hookrightarrow}_{q-1}, u_{q-1}\rangle, a_{q-1}\rangle \oplus g(\langle s_q\rangle, u^{\hookrightarrow}_q, a_{q-1}) \\
&= \tau^G_{0:q-1} \oplus \langle u_{q-1}, \langle s_{q-1}, u^{\hookrightarrow}_{q-1}, u_{q-1}\rangle, a_{q-1}, \langle s_q, u^{\hookrightarrow}_q, a_{q-1}\rangle\rangle \\
&= \tau^G.
\end{aligned}
$$

So if f is surjective for paths of arbitrary length $q-1 \in \mathbb{N}$, f is surjective for paths of length q. Hence, by induction, f is surjective.

f is injective and surjective, hence f is a bijection. □

Corollary 5 (Nature first bijections between histories)**.** *Let* $g^h \colon H^{M,\ltimes} \times A \to H^{G,\ltimes}$ *defined by:*

$$
\begin{aligned}
g^h(\langle z^{\mathfrak{a}}_\bullet, z^{\mathfrak{n}}_\bullet, z_\circ\rangle, a') &= \langle\langle z^{\mathfrak{n}}_\bullet, z_\circ, a'\rangle, \langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle\rangle. \\
g^h(\langle z^{\mathfrak{a}}_\bullet, z^{\mathfrak{n}}_\bullet, z_\circ, a, u\rangle, a') &= \langle\langle z^{\mathfrak{n}}_\bullet, z_\circ, a'\rangle, \langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle, u, \langle z^{\mathfrak{n}}_\bullet, z_\circ, \bot\rangle, \langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle, a\rangle. \\
g^h(\langle z^{\mathfrak{a}}_\bullet, z^{\mathfrak{n}}_\bullet, z_\circ, a, u\rangle \oplus h', a') &= g^h(\langle z^{\mathfrak{a}}_\bullet, z^{\mathfrak{n}}_\bullet, z_\circ, a, u\rangle, \bot) \oplus g^h(h', a).
\end{aligned}
$$

Let $f^h \colon H^M \to H^G$ *defined by:*

$$
\begin{aligned}
f^h(\langle z^{\mathfrak{a}}_\bullet, z^{\mathfrak{n}}_\bullet, z_\circ\rangle) &= \langle\langle z^{\mathfrak{n}}_\bullet, z_\circ, \bot\rangle, \langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle\rangle. \\
f^h(\langle z^{\mathfrak{a}}_\bullet, z^{\mathfrak{n}}_\bullet, z_\circ, a, u\rangle) &= \langle\langle z^{\mathfrak{n}}_\bullet, z_\circ, \bot\rangle, \langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle, u, \langle z^{\mathfrak{n}}_\bullet, z_\circ, \bot\rangle, \langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle, a\rangle. \\
f^h(\langle z^{\mathfrak{a}}_\bullet, z^{\mathfrak{n}}_\bullet, z_\circ, a, u\rangle \oplus h') &= f^h(\langle z^{\mathfrak{a}}_\bullet, z^{\mathfrak{n}}_\bullet, z_\circ, a, u\rangle) \oplus g^h(h', a).
\end{aligned}
$$

Let $g^{\mathfrak{a},h} \colon H^{\mathfrak{a},M,\ltimes} \to H^{\mathfrak{a},G,\ltimes}$ *defined by:*

$$
\begin{aligned}
g^{\mathfrak{a},h}(\langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle) &= \langle\langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle\rangle. \\
g^{\mathfrak{a},h}(\langle z^{\mathfrak{a}}_\bullet, z_\circ, a\rangle) &= \langle\langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle, \langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle, a\rangle. \\
g^{\mathfrak{a},h}(\langle z^{\mathfrak{a}}_\bullet, z_\circ, a, h'\rangle) &= g^{\mathfrak{a},h}(\langle z^{\mathfrak{a}}_\bullet, z_\circ, a\rangle) \oplus g^{\mathfrak{a},h}(h').
\end{aligned}
$$

Let $f^{\mathfrak{a},h} \colon H^{\mathfrak{a},M} \to H^{\mathfrak{a},G}$ *defined by:*

$$
\begin{aligned}
f^{\mathfrak{a},h}(\langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle) &= \langle\langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle\rangle. \\
f^{\mathfrak{a},h}(\langle z^{\mathfrak{a}}_\bullet, z_\circ, a\rangle) &= \langle\langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle, \langle z^{\mathfrak{a}}_\bullet, z_\circ\rangle, a\rangle. \\
f^{\mathfrak{a},h}(\langle z^{\mathfrak{a}}_\bullet, z_\circ, a, h'\rangle) &= f^{\mathfrak{a},h}(\langle z^{\mathfrak{a}}_\bullet, z_\circ, a\rangle) \oplus g^{\mathfrak{a},h}(h').
\end{aligned}
$$

$f^{\mathfrak{a},h}$ *is a bijection.*

Let $g^{\mathfrak{n},h} \colon H^{\mathfrak{n},M,\ltimes} \times A \to H^{\mathfrak{n},G,\ltimes}$ *defined by:*

$$
\begin{aligned}
g^{\mathfrak{n},h}(\langle z^{\mathfrak{n}}_\bullet, z_\circ\rangle, a') &= \langle\langle z^{\mathfrak{n}}_\bullet, z_\circ, a'\rangle\rangle. \\
g^{\mathfrak{n},h}(\langle z^{\mathfrak{n}}_\bullet, z_\circ, a, u\rangle, a') &= \langle\langle z^{\mathfrak{n}}_\bullet, z_\circ, a'\rangle, u, \langle z^{\mathfrak{n}}_\bullet, z_\circ, \bot\rangle\rangle. \\
g^{\mathfrak{n},h}(\langle z^{\mathfrak{n}}_\bullet, z_\circ, a, u, h'\rangle, a') &= g^{\mathfrak{n},h}(\langle z^{\mathfrak{n}}_\bullet, z_\circ, a, u\rangle, a') \oplus g^{\mathfrak{n},h}(h', a).
\end{aligned}
$$

Let $f^{\mathfrak{n},h} \colon H^{\mathfrak{n},M} \to H^{\mathfrak{n},G}$ *defined by:*

$$f^{\mathfrak{n},h}(\langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}\rangle) = \langle\langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, \bot\rangle\rangle.$$
$$f^{\mathfrak{n},h}(\langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a, u\rangle) = \langle\langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, \bot\rangle, u, \langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, \bot\rangle\rangle.$$
$$f^{\mathfrak{n},h}(\langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a, u, h'\rangle) = f^{\mathfrak{n},h}(\langle z^{\mathfrak{n}}_{\bullet}, z_{\circ}, a, u\rangle) \oplus g^{\mathfrak{n},h}(h', a).$$

$f^{\mathfrak{n},h}$ *is a bijection.*

The bijection between stochastic nature policies no longer involves an extra state observation, whereas the bijection between stochastic agent policy now does. Note that the extra agent state observation for the inverse function for the agent policy bijection can be derived from the history input, as this observation is the same as the last nature state observation contained in that history.

Corollary 6 (Bijection between policies)**.** *Let* $f^{\pi} \colon \Pi^{M} \to \Pi^{G}$ *defined by:*

$$f^{\pi}(\pi^{M})(h^{\mathfrak{a},G}, \langle z^{\mathfrak{a}}_{\bullet}, z_{\circ}\rangle) = \pi^{M}((f^{\mathfrak{a},h})^{-1}(h^{\mathfrak{a},G})),$$

then f^{π} *is a bijection.*

Let $f^{\theta} \colon \Theta^{M} \to \Theta^{G}$ *defined by:*

$$f^{\theta}(\theta^{M})(h^{\mathfrak{n},G}) = \theta^{M}((f^{\mathfrak{n},h})^{-1}(h^{\mathfrak{n},G})),$$

then f^{θ} *is a bijection.*

G Nash Equilibrium

This appendix contains all the proofs required to show the existence of a Nash equilibrium in our POSGs, *i.e.*, Theorem 3, restated below.

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:*

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

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

Throughout this appendix, we use the RPOMDP histories, paths, and policies, as these require simpler notation. We refer to Appendix F for the bijections between the RPOMDP and POSG paths, histories, and policies.

G.1 Sufficient Statistic

Where in POMDPs the history of the agent is enough to reason optimally, this is not the case for RPODMPs and their equivalent POSGs. Apart from their own history, the players must also consider all possible histories of the other player.

We adjust the notion of occupancy state used in [Delage *et al.*, 2023] to work with the infinite nature action space of our RPOMDP and equivalent POSGs. As mentioned in the introduction of this appendix, we use the RPOMDP notation. Given $\pi_{0:t-1} \in \Pi_{0:t-1}, \theta_{0:t-1} \in \Theta_{0:t-1}$, [Delage *et al.*, 2023] defines the occupancy state $\sigma_{\{\pi,\theta\}_{0:t-1}}$ as the probability distribution over all joint histories given agent and nature policies $\pi_{0:t-1}$ and $\theta_{0:t-1}$.

$$\forall h_t \in H_t : \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t) = \Pr(h_t \mid \pi_{0:t-1}, \theta_{0:t-1}).$$

$$\sum_{h_t \in H_t} \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t) = 1.$$

In the original definition of the occupancy state, nature's action space is finite. This occupancy state is a sufficient statistic for computing the next occupancy state and the expected reward at time t given the next π_t and θ_t in their POSG models.

However, as we deal with an infinite action space, we must make adjustments to ensure the subset of joint histories for each occupancy state is finite. Therefore, we keep track of nature's policy to be able to generate the finite subset of joint histories that the corresponding occupancy states have a distribution over.

Our version of the occupancy state extends the original occupancy state with the corresponding nature policy. Given $\pi_{0:t-1} \in \Pi_{0:t-1}, \theta_{0:t-1} \in \Theta_{0:t-1}$:

$$\text{OS}_{\{\pi,\theta\}_{0:t-1}} \stackrel{\text{def}}{=} \langle \sigma_{\{\pi,\theta\}_{0:t-1}}, \theta_{0:t-1} \rangle.$$

$$\forall h_t \in H_t : \sigma_{\{\pi,\theta\}_{0:t-1}}\theta_{0:t-1}(h_t) = \Pr(h_t \mid \pi_{0:t-1}, \theta_{0:t-1}).$$

$$\sum_{h_t \in H_t} \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t) = 1.$$

We show that the occupancy state $\text{OS}_{\{\pi,\theta\}_{0:t-1}}$ together with agent and nature policies π_t, θ_t at time t, is a sufficient statistic for computing the next occupancy state $\text{OS}_{\{\pi,\theta\}_{0:t}}$ and the expected reward $\mathsf{R}(\text{OS}_{\{\pi,\theta\}_{0:t-1}}, \pi_t, \theta_t) = \mathbb{E}[r_t \mid \pi_{0:t-1}, \pi_t, \theta_{0:t-1}, \theta_t]$. Note that these proofs are based on the proofs in Appendix B of [Delage *et al.*, 2023].

$$\text{OS}_{\{\pi,\theta\}_{0:t}}(h_t \oplus \langle a_t, u_t, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \rangle) \stackrel{\text{def}}{=} \langle \sigma_{\{\pi,\theta\}_{0:t}}(h_t \oplus \langle a_t, u_t, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \rangle), \theta_{0:t} \rangle.$$

Where:

$$\theta_{0:t} \stackrel{\text{def}}{=} \theta_{0:t-1} \oplus \theta_t.$$

$$\sigma_{\{\pi,\theta\}_{0:t}}(h_t \oplus \langle a_t, u_t, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \rangle) \stackrel{\text{def}}{=} \Pr(h_t, a_t, u_t, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid \pi_{0:t}, \theta_{0:t})$$

$$= \sum_{s,s' \in S} \Pr(h_t, a_t, u_t, z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ, s, s' \mid \pi_{0:t}, \theta_{0:t})$$

$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid h_t, a_t, u_t, s, s', \pi_{0:t}, \theta_{0:t}) \Pr(h_t, a_t, u_t, s, s' \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of an observation only depends on the state:

$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(h_t, a_t, u_t, s, s' \mid \pi_{0:t}, \theta_{0:t})$$

$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(s' \mid h_t, a_t, u_t, s, \pi_{0:t}, \theta_{0:t}) \Pr(h_t, a_t, u_t, s \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of reaching a state only depends on the previous state and the agent and nature actions:

$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(h_t, a_t, u_t, s \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, s, \pi_{0:t}, \theta_{0:t}) \Pr(h_t, a_t, s \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of a nature action only depends on nature's policy at time t, the history, and the agent action at time t:

$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, \theta_t) \Pr(h_t, a_t, s \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, \theta_t) \Pr(a_t \mid h_t, s, \pi_{0:t}, \theta_{0:t}) \Pr(h_t, s \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of an agent action only depends on the agent's policy at time t, and the history:

$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, \theta_t) \Pr(a_t \mid h_t, \pi_t) \Pr(h_t, s \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, \theta_t) \Pr(a_t \mid h_t, \pi_t) \Pr(s \mid h_t, \pi_{0:t}, \theta_{0:t}) \Pr(h_t \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of being in a state can be computed via the belief generated by the joint history (see Appendix A.2):

$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, \theta_t) \Pr(a_t \mid h_t, \pi_t) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of a history at time t does not depend on actions of time t:

$$= \sum_{s,s' \in S} \Pr(z_\bullet^{\mathfrak{a}}, z_\bullet^{\mathfrak{n}}, z_\circ \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, \theta_t) \Pr(a_t \mid h_t, \pi_t) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t-1}, \theta_{0:t-1})$$
$$= \sum_{s,s' \in S} O_\bullet^{\mathfrak{a}}(s', z_\bullet^{\mathfrak{a}}) O_\bullet^{\mathfrak{n}}(s', z_\bullet^{\mathfrak{n}}) O_\circ(s', z_\circ) \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, \theta_t) \Pr(a_t \mid h_t, \pi_t) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s,s' \in S} O_\bullet^{\mathfrak{a}}(s', z_\bullet^{\mathfrak{a}}) O_\bullet^{\mathfrak{n}}(s', z_\bullet^{\mathfrak{n}}) O_\circ(s', z_\circ) \boldsymbol{T}(u_t)(s, a_t)(s') \Pr(u_t \mid h_t, a_t, \theta_t) \Pr(a_t \mid h_t, \pi_t) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s,s' \in S} O_\bullet^{\mathfrak{a}}(s', z_\bullet^{\mathfrak{a}}) O_\bullet^{\mathfrak{n}}(s', z_\bullet^{\mathfrak{n}}) O_\circ(s', z_\circ) \boldsymbol{T}(u_t)(s, a_t)(s') \theta_t(h_t^{\mathfrak{n}}, a_t)(u_t) \Pr(a_t \mid h_t, \pi_t) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s,s' \in S} O_\bullet^{\mathfrak{a}}(s', z_\bullet^{\mathfrak{a}}) O_\bullet^{\mathfrak{n}}(s', z_\bullet^{\mathfrak{n}}) O_\circ(s', z_\circ) \boldsymbol{T}(u_t)(s, a_t)(s') \theta_t(h_t^{\mathfrak{n}}, a_t)(u_t) \pi_t(h_t^{\mathfrak{a}})(a_t) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s,s' \in S} O_\bullet^{\mathfrak{a}}(s', z_\bullet^{\mathfrak{a}}) O_\bullet^{\mathfrak{n}}(s', z_\bullet^{\mathfrak{n}}) O_\circ(s', z_\circ) \boldsymbol{T}(u_t)(s, a_t)(s') \theta_t(h_t^{\mathfrak{n}}, a_t)(u_t) \pi_t(h_t^{\mathfrak{a}})(a_t) b(s, h_t) \Pr(h_t \mid \pi_{0:t}, \theta_{0:t}).$$

Where $b(s, h_t)$ is the belief computed by t belief updates given the joint history h_t (see Appendix A.2).

$$= \sum_{s,s' \in S} O_\bullet^{\mathfrak{a}}(s', z_\bullet^{\mathfrak{a}}) O_\bullet^{\mathfrak{n}}(s', z_\bullet^{\mathfrak{n}}) O_\circ(s', z_\circ) \boldsymbol{T}(u_t)(s, a_t)(s') \theta_t(h_t^{\mathfrak{n}}, a_t)(u_t) \pi_t(h_t^{\mathfrak{a}})(a_t) b(s, h_t) \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t).$$

This shows that we can compute the successor occupancy state using only the previous occupancy state $\mathrm{OS}_{\{\pi,\theta\}_{0:t-1}} = \langle \sigma_{\{\pi,\theta\}_{0:t-1}}, \theta_{0:t-1} \rangle$ and policies π_t, θ_t at time t. Note that we can use the nature policy $\theta_{0:t}$ to generate the finite subset of relevant histories $\mathsf{rel}(\theta_{0:t}) \subset H$ that possibly have a non-zero probability. We can then select the relevant histories of t time steps, denoted as $\mathsf{rel}(\theta_{0:t})_t \subset H_t$, to compute the occupancy states. See Appendix A.2 for details on the set of relevant histories. Next, we look at the expected reward.

$$\mathbb{E}[r_t \mid \pi_{0:t}, \theta_{0:t}] = \sum_{s \in S} \sum_{a \in A} R(s, a) \Pr(s, a \mid \pi_{0:t}, \theta_{0:t})$$

$$= \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \Pr(s, a, h_t \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \Pr(a \mid s, h_t, \pi_{0:t}, \theta_{0:t}) \Pr(s, h_t \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of an agent action only depends on the agents's policy at time t, and the history:

$$= \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \Pr(a \mid h_t, \pi_t) \Pr(s, h_t \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \Pr(a \mid h_t, \pi_t) \Pr(s \mid h_t, \pi_{0:t}, \theta_{0:t}) \Pr(h_t \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of being in a state can be computed via the belief generated by the joint history (see Appendix A.2):

$$= \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \Pr(a \mid h_t, \pi_t) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of a history at time t does not depend on actions of time t:

$$= \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \Pr(a \mid h_t, \pi_t) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t-1}, \theta_{0:t-1})$$
$$= \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \pi_t(h_t^{\mathtt{a}})(a) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t-1}, \theta_{0:t-1})$$
$$= \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \pi_t(h_t^{\mathtt{a}})(a) b(s, h_t) \Pr(h_t \mid \pi_{0:t-1}, \theta_{0:t-1})$$
$$= \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \pi_t(h_t^{\mathtt{a}})(a) b(s, h_t) \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t).$$

This shows that we can compute the expected reward at time t using only the previous occupancy state $\mathrm{OS}_{\{\pi,\theta\}_{0:t-1}} = (\sigma_{\{\pi,\theta\}_{0:t-1}}, \theta_{0:t-1})$ and policy π_t at time t.

Remark 5. The equivalent formulation of the occupancy state using POSG notation is as follows:

$$\mathrm{OS}_{\{\pi,\theta\}_{0:t}}(h_t \oplus \langle\langle z_\bullet^{\mathtt{a}}, z_\circ\rangle, \langle z_\bullet^{\mathtt{n}}, z_\circ, \bot\rangle, a_t, \langle z_\bullet^{\mathtt{a}}, z_\circ\rangle, \langle z_\bullet^{\mathtt{n}}, z_\circ, a_t\rangle, u_t\rangle)$$
$$\stackrel{\text{def}}{=} \langle \sigma_{\{\pi,\theta\}_{0:t}}(h_t \oplus \langle\langle z_\bullet^{\mathtt{a}}, z_\circ\rangle, \langle z_\bullet^{\mathtt{n}}, z_\circ, \bot\rangle, a_t, \langle z_\bullet^{\mathtt{a}}, z_\circ\rangle, \langle z_\bullet^{\mathtt{n}}, z_\circ, a_t\rangle, u_t\rangle), \theta_{0:t}\rangle.$$

Where:

$$\theta_{0:t} \stackrel{\text{def}}{=} \theta_{0:t-1} \oplus \theta_t.$$
$$\sigma_{\{\pi,\theta\}_{0:t}}(h_t \oplus \langle\langle z_\bullet^{\mathtt{a}}, z_\circ\rangle, \langle z_\bullet^{\mathtt{n}}, z_\circ, \bot\rangle, a_t, \langle z_\bullet^{\mathtt{a}}, z_\circ\rangle, \langle z_\bullet^{\mathtt{n}}, z_\circ, a_t\rangle, u_t\rangle) \stackrel{\text{def}}{=} \Pr(h_t, a_t, u_t, z_\bullet^{\mathtt{a}}, z_\bullet^{\mathtt{n}}, z_\circ \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s \in \mathcal{S}^{\mathtt{a}}} \sum_{s' \in \mathcal{S}^{\mathtt{a}}} \mathcal{O}^{\mathtt{a}}(s', \langle z_\bullet^{\mathtt{a}}, z_\circ\rangle) \mathcal{O}^{\mathtt{n}}(s', \langle z_\bullet^{\mathtt{n}}, z_\circ\rangle) \mathcal{T}^{\mathtt{n}}(\mathcal{T}^{\mathtt{a}}(s, a_t), u_t, s') \theta_t(h_t^{\mathtt{n}}, a_t)(u_t) \pi_t(h_t^{\mathtt{a}})(a_t) b(s, h_t) \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t).$$

We can compute the expected reward as follows:

$$\mathbb{E}[r_t \mid \pi_{0:t}, \theta_{0:t}] = \sum_{s \in \mathcal{S}^{\mathtt{a}}} \sum_{a \in \mathcal{A}^{\mathtt{a}}} \mathcal{R}(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \pi_t(h_t^{\mathtt{a}})(a) b(s, h_t) \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t).$$

This formulation again shows that the occupancy state $\mathrm{OS}_{\{\pi,\theta\}_{0:t-1}}$ together with agent and nature policies π_t, θ_t at time t, is a sufficient statistic for computing the next occupancy state $\mathrm{OS}_{\{\pi,\theta\}_{0:t}}$ and the expected reward $\mathsf{R}(\mathrm{OS}_{\{\pi,\theta\}_{0:t-1}}, \pi_t, \theta_t)$.

G.2 Occupancy Game

Given the occupancy state, we can define a non-observable, non-stochastic game: a zero-sum occupancy game (OG) [Delage *et al.*, 2023]. As the occupancy state is a sufficient statistic, computing a Nash equilibrium in this OG is equivalent to a Nash equilibrium in our original RPOMDP.

Definition 19 (Occupancy game)**.** *Given an RPOMDP* $\langle S, A, \boldsymbol{T}, R, Z_\bullet^{\mathtt{a}}, Z_\bullet^{\mathtt{n}}, Z_\circ, O_\bullet^{\mathtt{a}}, O_\bullet^{\mathtt{n}}, O_\circ, \mathsf{stick}, \mathfrak{a}\rangle$*, and a horizon* $K \in \mathbb{N}$*, we define the occupancy game as a tuple* $(\mathsf{S}^{\mathtt{a}}, \mathsf{S}^{\mathtt{n}}, \mathsf{A}^{\mathtt{a}}, \mathsf{A}^{\mathtt{n}}, \mathsf{T}, \mathsf{R})$ *where the sets of states and actions are defined as follows:* $\mathsf{S}^{\mathtt{a}} = \bigcup_{t=0}^{K-1} \bigcup_{\pi_{0:t} \in \Pi_{0:t}} \bigcup_{\theta_{0:t} \in \Theta_{0:t}} OS_{\{\pi,\theta\}_{0:t}}$ *is the infinite set of agent states, and* $\mathsf{S}^{\mathtt{n}} = \bigcup_{t=0}^{K-1} (\bigcup_{\pi_{0:t} \in \Pi_{0:t}} \bigcup_{\theta_{0:t} \in \Theta_{0:t}} OS_{\{\pi,\theta\}_{0:t}} \times \Pi_{t+1})$ *the infinite set of nature states;* $\mathsf{A}^{\mathtt{a}} = \bigcup_{t=0}^{K-1} \Pi_t$ *is the infinite set of agent actions, and* $\mathsf{A}^{\mathtt{n}} = \bigcup_{t=0}^{K-1} \Theta_t$ *the infinite set of nature actions; The transition and reward functions are then defined as:*

- $\mathsf{T} = \mathsf{T}^{\mathtt{a}} \cup \mathsf{T}^{\mathtt{n}}$*, the transition function, where:*
  - $\mathsf{T}^{\mathtt{a}} \colon \mathsf{S}^{\mathtt{a}} \times \mathsf{A}^{\mathtt{a}} \hookrightarrow \mathsf{S}^{\mathtt{n}}$ *the agent's transition function.*
  - $\mathsf{T}^{\mathtt{n}} \colon \mathsf{S}^{\mathtt{n}} \times \mathsf{A}^{\mathtt{n}} \hookrightarrow \mathsf{S}^{\mathtt{a}}$ *nature's transition function.*
- $\mathsf{R} \colon \mathsf{S}^{\mathtt{a}} \times \mathsf{A}^{\mathtt{a}} \to \mathbb{R}$ *the reward function.*

Where:

- $\mathsf{R}(\langle \sigma_{\{\pi,\theta\}_{0:t}}, \theta_{0:t} \rangle, \pi_{t+1}) = \sum_{s \in S} \sum_{a \in A} R(s,a) \sum_{h_t \in \mathsf{rel}(\theta_{0:t-1})_t} \pi_t(h_t^{\mathtt{a}}, a) b(s, h_t) \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t)$.
- $\mathsf{T}^{\mathtt{a}}(\langle \sigma_{\{\pi,\theta\}_{0:t}}, \theta_{0:t} \rangle, \pi_{t+1}) = \langle \langle \sigma_{\{\pi,\theta\}_{0:t}}, \theta_{0:t} \rangle, \pi_{t+1} \rangle$.
- $\mathsf{T}^{\mathtt{n}}(\langle \langle \sigma_{\{\pi,\theta\}_{0:t}}, \theta_{0:t} \rangle, \pi_{t+1} \rangle, \theta_{t+1}) = \langle \sigma_{\{\pi,\theta\}_{0:t+1}}, \theta_{0:t+1} \rangle$*, where:*
  - $\theta_{0:t+1} = \theta_{0:t} \oplus \theta_{t+1}$.
  - $\forall h_{t+1} \in \mathsf{rel}(\theta_{0:t})_{t+1}, \forall a \in \mathcal{A}^{\mathtt{a}}, \forall u \in \mathcal{A}^{\mathtt{n}}, \forall z_\bullet^{\mathtt{a}}, z_\bullet^{\mathtt{n}}, z_\circ \in Z_\bullet^{\mathtt{a}} \times Z_\bullet^{\mathtt{n}} \times Z_\circ, \sigma_{\{\pi,\theta\}_{0:t+1}}(\langle h_{t+1}, a, u, z_\bullet^{\mathtt{a}}, z_\bullet^{\mathtt{n}}, z_\circ \rangle) =$

$$\sum_{s,s' \in S} O_\bullet^{\mathtt{a}}(s', z_\bullet^{\mathtt{a}}) O_\bullet^{\mathtt{n}}(s', z_\bullet^{\mathtt{n}}) O_\circ(s', z_\circ) \boldsymbol{T}(u_t)(s, a_t)(s') \theta_t(h_t^{\mathtt{n}}, a_t)(u_t) \pi_t(h_t^{\mathtt{a}})(a_t) b(s, h_t) \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t).$$

Where $b(s, h_t)$ is the belief computed by t belief updates given the joint history h_t (see Appendix A.2). For deriving the reward and transition functions, see Appendix G.1.

G.3 Mixed policies

As shown in Appendix G.1, the occupancy state is a sufficient statistic for the POSG and, hence, for the RPOMDP. In appendix G.4, we show that the occupancy game has a Nash equilibrium for finite horizon reward maximization, and an optimal policy for the agent exists. To prove that this Nash equilibrium exists, we need to reason with mixed policies instead of stochastic policies. In this section, we prove that reasoning with the set of mixed agent and nature policies results in the same set of distributions over paths, and hence the same possible values, as reasoning with the set of stochastic policies.

Concretely, we need to show that for every stochastic policy, there exists a mixed policy that behaves equivalently, meaning it results in the same distribution over paths. We focus on the RPOMDP policies. The same results follow for the POSG policies using the bijections from Appendix F.

Theorem 5 (Existence of equivalent mixed policy)**.** *Let* $\mu^{\pi,\theta} \in \Delta(Paths^M)$ *be the probability distribution over paths in the RPOMDP resulting from executing agent policy* π *and nature policy* θ*. Then:*

$$\forall \pi \in \Pi, \exists \pi^{mix} \in \Pi^{mix}, \forall \theta \in \Theta \cup \Theta^{mix}. \mu^{\pi,\theta} = \mu^{\pi^{mix},\theta},$$

$$\forall \theta \in \Theta, \exists \theta^{mix} \in \Theta^{mix}, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta} = \mu^{\pi,\theta^{mix}}.$$

Before proving the theorem above, we consider the other key results that follow. We immediately get the following corollary from Theorem 5.

Corollary 7 (stochastic policies $\subseteq$ mixed policies)**.** *Let* $\mu^{\pi,\theta} \in \Delta(Paths^M)$ *be the probability distribution over paths in the RPOMDP resulting from executing agent policy* π *and nature policy* θ*. Then we have the following:*

$$\{\mu^{\pi,\theta} \mid \pi \in \Pi, \theta \in \Theta\} \subseteq \{\mu^{\pi^{mix},\theta^{mix}} \mid \pi^{mix} \in \Pi^{mix}, \theta^{mix} \in \Theta^{mix}\}.$$

We also need to show that for every mixed policy, there exists a stochastic policy that behaves equivalently. This means that there are no new behaviors, and consequently no new values, introduced by looking at the set of mixed policies.

Theorem 6 (Existence of equivalent stochastic policy)**.** *Let* $\mu^{\pi,\theta} \in \Delta(Paths^M)$ *be the probability distribution over paths in the RPOMDP resulting from executing agent policy* π *and nature policy* θ*. Then:*

$$\forall \pi^{mix} \in \Pi^{mix}, \exists \pi \in \Pi, \forall \theta \in \Theta \cup \Theta^{mix}. \mu^{\pi^{mix},\theta} = \mu^{\pi,\theta},$$

$$\forall \theta^{mix} \in \Theta^{mix}, \exists \theta \in \Theta, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta^{mix}} = \mu^{\pi,\theta}.$$

Theorem 6 comes with the following corollary.

Corollary 8 (stochastic policies $\supseteq$ mixed policies)**.** *Let $\mu^{\pi,\theta} \in \Delta(Paths^M)$ be the probability distribution over paths in the RPOMDP resulting from executing agent policy π and nature policy θ. Then we have the following:*

$$\{\mu^{\pi,\theta} \mid \pi \in \Pi, \theta \in \Theta\} \supseteq \{\mu^{\pi^{mix},\theta^{mix}} \mid \pi^{mix} \in \Pi^{mix}, \theta^{mix} \in \Theta^{mix}\}.$$

By combining Corollaries 7 and 8, it follows that the sets of mixed policies give exactly the same sets of distributions over paths in our original RPOMDP as the sets of stochastic policies do.

Corollary 9 (Equivalent set of mixed policies)**.** *Let $\mu^{\pi,\theta} \in \Delta(Paths^M)$ be the probability distribution over paths in the RPOMDP resulting from executing agent policy π and nature policy θ. Then we have the following:*

$$\{\mu^{\pi,\theta} \mid \pi \in \Pi, \theta \in \Theta\} = \{\mu^{\pi^{mix},\theta^{mix}} \mid \pi^{mix} \in \Pi^{mix}, \theta^{mix} \in \Theta^{mix}\}.$$

It follows that the sets of possible values, i.e.*, the subset of $\mathbb{R}$ the value function can attain under all stochastic policies and all mixed policies, is the same:*

$$\{V^{\pi,\theta} \mid \pi \in \Pi, \theta \in \Theta\} = \{V^{\pi^{mix},\theta^{mix}} \mid \pi^{mix} \in \Pi^{mix}, \theta^{mix} \in \Theta^{mix}\}.$$

Below, we first define some definitions and lemmas and then prove Theorems 5 and 6. We focus on the nature policies, as these require dealing with an infinite action space and, therefore, with an infinite set of deterministic policies. The proofs for the agent policies follow the same steps.

Additional definitions and lemmas

We begin by defining the set of relevant histories given a policy. The actions chosen for histories outside the relevant history set do not influence the results of a game since the policy never reaches them.

Using the relevant histories, we can define the set of relevant deterministic policies given a history.

Definition 20 (Relevant deterministic policies)**.** *Given a history $h^{\mathtt{n}} \in H^{\mathtt{n}}$, we define the set of relevant deterministic policies $\Theta^{det,h}$, containing all deterministic policies that could have generated the current history.*

$$\Theta^{det,h^{\mathtt{n}}} = \{\theta^{det} \in \Theta^{det} \mid h^{\mathtt{n}} \in \mathsf{rel}^{\mathtt{n}}(\theta^{det})\}.$$

We define a helper function $\eta^{\pi,\theta} \colon \Delta(\text{Paths}^M)$ to compute the probability over paths for all stochastic and deterministic policies $\pi \in \Pi$ and $\theta \in \Theta$:

$$\eta^{\pi,\theta}(\langle s_I \rangle) = 1,$$
$$\eta^{\pi,\theta}(\tau' \oplus \langle s, a, u, s' \rangle) = \eta^{\pi,\theta}(\tau' \oplus \langle s \rangle) \cdot \pi(O^{\mathtt{a},M}(\tau' \oplus \langle s \rangle))(a) \cdot \theta(O^{\mathtt{n},M}(\tau' \oplus \langle s \rangle), a)(u) \cdot \boldsymbol{T}(u)(s, a, s').$$

We now prove the following lemmas about the probability distribution over paths for deterministic policies.

Given a path, if a deterministic policy is not relevant for generating the history of that path, then the probability of reaching that path with the deterministic policy is zero.

Lemma 5 (Zero probability of non-relevant paths)**.** *Given a path τ and policies $\pi \in \Pi$, $\theta^{det} \in \Theta^{det}$, we have that:*

$$\theta^{det} \notin \Theta^{det,O^{\mathtt{n},M}(\tau)} \implies \eta^{\pi,\theta^{det}}(\tau) = 0.$$

Proof. Take arbitrary path τ and policies $\pi \in \Pi$, $\theta^{det} \in \Theta^{det}$.

$$\theta^{det} \notin \Theta^{det,O^{\mathtt{n},M}(\tau)} \iff O^{\mathtt{n},M}(\tau) \notin \mathsf{rel}^{\mathtt{n}}(\theta^{det}).$$

By definition, we know that $O^{\mathtt{n},M}(\langle s_I \rangle) \in \mathsf{rel}^{\mathtt{n}}(\theta^{det})$. Furthermore, we know that:

$$O^{\mathtt{n},M}(\tau' \oplus \langle a', u', s \rangle) \notin \mathsf{rel}^{\mathtt{n}}(\theta^{det}) \iff \theta^{det}(O^{\mathtt{n},M}(\tau'), a') \neq u' \vee O^{\mathtt{n},M}(\tau') \notin \mathsf{rel}^{\mathtt{n}}(\theta^{det}).$$

Since $O^{\mathtt{n},M}(\langle s_I \rangle) \in \mathsf{rel}^{\mathtt{n}}(\theta^{det})$, we will eventually reach a prefix of τ, for which the condition is violated. So then we get:

$$\begin{aligned}
O^{\mathtt{n},M}(\tau) \notin \mathsf{rel}^{\mathtt{n}}(\theta^{det}) &\iff \exists \tau''.\, \tau'' \oplus \langle a'', u'', s', \ldots \rangle = \tau \wedge \theta^{det}(O^{\mathtt{n},M}(\tau''), a'') \neq u'' \\
&\iff \exists \tau''.\, \tau'' \oplus \langle a'', u'', s', \ldots \rangle = \tau \wedge \theta^{det}(O^{\mathtt{n},M}(\tau''), a'')(u'') = 0 \\
&\implies \exists \tau''.\, \tau'' \oplus \langle a'', u'', s', \ldots \rangle = \tau \wedge \eta^{\pi,\theta^{det}}(\tau'' \oplus \langle a'', u'', s' \rangle) = 0 \\
&\implies \eta^{\pi,\theta^{det}}(\tau) = 0.
\end{aligned}$$

Hence:

$$\theta^{det} \notin \Theta^{det,O^{\mathtt{n},M}(\tau)} \implies \eta^{\pi,\theta^{det}}(\tau) = 0.$$

□

The next lemma states that, given a path, if two deterministic policies are both relevant for generating the history of that path, then the probability of reaching that path is the same for both policies.

Lemma 6 (Constant probability of paths for relevant deterministic policies)**.** *Given a path τ and policy $\pi \in \Pi$, we have that:*

$$\forall \theta^{det}, \theta^{det'} \in \Theta^{det, O^{\mathfrak{n},M}(\tau)}. \eta^{\pi,\theta^{det}}(\tau) = \eta^{\pi,\theta^{det'}}(\tau).$$

Proof. Take arbitrary path τ, policy $\pi \in \Pi$ and $\theta^{det}, \theta^{det'} \in \Theta^{det, O^{\mathfrak{n},M}(\tau)}$. We show $\eta^{\pi,\theta^{det}}(\tau) = \eta^{\pi,\theta^{det'}}(\tau)$ by induction on the length of τ.

Assume $|\tau| = 0$. Then $\tau = \langle s_I \rangle$. Then:

$$\eta^{\pi,\theta^{det}}(\langle s_I \rangle) = 1 = \eta^{\pi,\theta^{det'}}(\langle s_I \rangle).$$

So for paths τ of length 0, we know that $\eta^{\pi,\theta^{det}}(\tau) = \eta^{\pi,\theta^{det'}}(\tau)$.

Now assume we know, given $q \in \mathbb{N}, q \geq 1$, that:

$$\forall \tau \in \text{Paths}^M. |\tau| = q - 1 \implies \eta^{\pi,\theta^{det}}(\tau) = \eta^{\pi,\theta^{det'}}(\tau).$$

Take arbitrary $\tau \in \text{Paths}^M$ with horizon length $|\tau| = q$. Then we have:

$$\tau = \tau_{0:q-1} \oplus \langle a_{q-1}, u_{q-1}, s_q \rangle = \tau_{0:q-2} \oplus \langle a_{q-2}, u_{q-2}, s_{q-1}, a_{q-1}, u_{q-1}, s_q \rangle.$$

Then $\tau_{0:q-1} \in \text{Paths}^M$ and $|\tau_{0:q-1}| = q - 1$. By assumption, we get that:

$$\eta^{\pi,\theta^{det}}(\tau_{0:q-1}) = \eta^{\pi,\theta^{det'}}(\tau_{0:q-1}).$$

Additionally, we know that:

$$\begin{aligned}
\theta^{det} \in \Theta^{det, O^{\mathfrak{n},M}(\tau)} &\iff O^{\mathfrak{n},M}(\tau) \in \mathsf{rel}^{\mathfrak{n}}(\theta^{det}) \\
&\iff \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1}) = u_{q-1} \wedge O^{\mathfrak{n},M}(\tau_{0:q-1}) \in \mathsf{rel}^{\mathfrak{n}}(\theta^{det}) \\
&\iff \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) = 1 \wedge O^{\mathfrak{n},M}(\tau_{0:q-1}) \in \mathsf{rel}^{\mathfrak{n}}(\theta^{det}).
\end{aligned}$$

So we have that:

$$\theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) = 1 = \theta^{det'}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}).$$

Finally, we get:

$$\begin{aligned}
\eta^{\pi,\theta^{det}}(\tau) &= \eta^{\pi,\theta^{det}}(\tau_{0:q-1}) \cdot \pi(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) \cdot \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \\
&= \eta^{\pi,\theta^{det'}}(\tau_{0:q-1}) \cdot \pi(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \theta^{det'}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) \cdot \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \\
&= \eta^{\pi,\theta^{det'}}(\tau).
\end{aligned}$$

So, if $\eta^{\pi,\theta^{det}}(\tau) = \eta^{\pi,\theta^{det'}}(\tau)$ holds for paths of arbitrary length $q - 1 \in \mathbb{N}$, then $\eta^{\pi,\theta^{det}}(\tau) = \eta^{\pi,\theta^{det'}}(\tau)$ holds for paths of length q. Hence, by induction, $\eta^{\pi,\theta^{det}}(\tau) = \eta^{\pi,\theta^{det'}}(\tau)$.

As $\theta^{det}, \theta^{det'} \in \Theta^{det, O^{\mathfrak{n},M}(\tau)}$ were arbitrarily chosen, we conclude that:

$$\forall \theta^{det}, \theta^{det'} \in \Theta^{det, O^{\mathfrak{n},M}(\tau)}. \eta^{\pi,\theta^{det}}(\tau) = \eta^{\pi,\theta^{det'}}(\tau).$$

□

Using our helper function η, we can define the four ways of computing the probability distribution over paths depending on the type of policies involved as follows:

(1) $\pi \in \Pi$ and $\theta \in \Theta$:

$$\mu^{\pi,\theta}(\tau) = \eta^{\pi,\theta}(\tau).$$

(2) $\pi \in \Pi$ and $\theta^{mix} \in \Theta^{mix}$:

$$\mu^{\pi,\theta^{mix}}(\tau) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi,\theta^{det}}(\tau).$$

(3) $\pi^{mix} \in \Pi^{mix}$ and $\theta \in \Theta$:

$$\mu^{\pi^{mix},\theta}(\tau) = \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \eta^{\pi^{det},\theta}(\tau).$$

(4) $\pi^{mix} \in \Pi^{mix}$ and $\theta^{mix} \in \Theta^{mix}$:

$$\begin{aligned}\mu^{\pi^{mix},\theta^{mix}}(\tau) &= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det},\theta^{det}}(\tau), \\ &= \sum_{\pi^{det} \in \Pi^{det}} \sum_{\theta^{det} \in \Theta^{det}} \pi^{mix}(\pi^{det}) \cdot \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det},\theta^{det}}(\tau).\end{aligned}$$

Recall that a deterministic policy can be interpreted as both a stochastic and a mixed policy using only Dirac distributions. The probabilities over paths generated by deterministic policies can, therefore, be computed using any of the above formulas.

Proof of Theorem 5

The standard way to define a mixed strategy given a stochastic strategy is to simply assign to each deterministic policy the product of the probabilities the stochastic policy assigns to the same choices [Kuhn, 1953]. The problem in our case, however, is that, due to the infinite number of nature policies, this leads to infinitely many deterministic policies having a non-zero probability in the resulting mixed policy.

To create a mixed policy with a finite number of deterministic policies with a non-zero probability, we define an equivalence class for deterministic policies that assign the same action for all histories that are relevant to the given stochastic policy.

Definition 21. *Given a stochastic policy $\theta \in \Theta$, we define the following equivalence relation $\sim_{\mathsf{rel}^{\mathtt{n}}(\theta)}$ between deterministic policies, which we call $\mathsf{rel}^{\mathtt{n}}(\theta)$-equivalent:*

$$\forall \theta^{det}, \theta^{det'} \in \Theta^{det}.\, \theta^{det} \sim_{\mathsf{rel}^{\mathtt{n}}(\theta)} \theta^{det'} \iff \forall h^{\mathtt{n}} \in \mathsf{rel}^{\mathtt{n}}(\theta), \forall a \in A.\, \theta^{det}(h^{\mathtt{n}}, a) = \theta^{det'}(h^{\mathtt{n}}, a).$$

The reflexivity, symmetry, and transitivity of the $\mathsf{rel}^{\mathtt{n}}(\theta)$-equivalence relation follow from the reflexivity, symmetry, and transitivity of the equality relation.

The $\mathsf{rel}^{\mathtt{n}}(\theta)$-equivalence relation provides us with $\mathsf{rel}^{\mathtt{n}}(\theta)$-equivalence classes $[\theta^{det}]_{\sim_{\mathsf{rel}^{\mathtt{n}}(\theta)}}$, which partition the set of deterministic policies. We select one member of each $\mathsf{rel}^{\mathtt{n}}(\theta)$-equivalence class to define a new set $\theta^{det,\sim_{\mathsf{rel}^{\mathtt{n}}(\theta)}}$ called the $\mathsf{rel}^{\mathtt{n}}(\theta)$-representation set. Note that this set is not the same as the quotient set $\Theta^{det}/\sim_{\mathsf{rel}^{\mathtt{n}}(\theta)}$, as the quotient set is a set of sets of deterministic policies, whereas the $\mathsf{rel}^{\mathtt{n}}(\theta)$-representation set is a set of deterministic policies. Clearly, $\Theta^{det}/\sim_{\mathsf{rel}^{\mathtt{n}}(\theta)} \subseteq \Theta^{det}$.

Using the $\mathsf{rel}^{\mathtt{n}}(\theta)$-representation set, we define a function $g\colon \Theta \to \Theta^{mix}$ with which we will construct our equivalent mixed policy. This function uses a similar construction as in [Kuhn, 1953] for the deterministic policies in the $\mathsf{rel}^{\mathtt{n}}(\theta)$-representation set but gives the rest of the deterministic policies a zero probability automatically.

$$g(\theta)(\theta^{det}) = \begin{cases} \prod_{h^{\mathtt{n}} \in \mathsf{rel}^{\mathtt{n}}(\theta), a \in A} \theta(h, a)(\theta^{det}(h^{\mathtt{n}}, a)) & \text{if } \theta^{det} \in \Theta^{det,\sim_{\mathsf{rel}^{\mathtt{n}}(\theta)}}, \\ 0 & \text{otherwise.} \end{cases}$$

We first show that g correctly maps to a mixed policy (Lemma 7) and then that this resulting policy results in the same distribution over paths in the RPOMDP given any agent policy (Lemma 8).

Lemma 7. *$g(\theta)$ is a mixed policy:*

$$\forall \theta \in \Theta.\, g(\theta) \in \Theta^{mix}.$$

Proof. Take arbitrary $\theta \in \Theta$. By construction, we have that $g(\theta)\colon \Theta^{det} \to [0, 1]$. Now to show that $g(\theta) \in \Theta^{mix}$, we must show two things: $g(\theta)$ assigns a non-zero probability to a finite number of deterministic policies (finitely randomizing) and $g(\theta)$ is a probability distribution, meaning the probabilities sum up to 1.

$$g(\theta) \text{ is finitely randomizing,} \tag{2}$$

$$\sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) = 1. \tag{3}$$

When proving Equation (2), we can restrict ourselves to $\theta^{det} \in \Theta^{det,\sim_{\mathsf{rel}^{\mathtt{n}}(\theta)}}$, as we assign a zero probability to all other deterministic policies. There can be infinitely many $\mathsf{rel}^{\mathtt{n}}(\theta)$-equivalence classes and, therefore, infinitely many elements of $\Theta^{det,\sim_{\mathsf{rel}^{\mathtt{n}}(\theta)}}$. However, we know that there is a finite number of nature histories in $\mathsf{rel}^{\mathtt{n}}(\theta)$ since θ is finitely randomizing, $Z^{\mathtt{n}}_{\bullet}$, $Z_{\circ}$, and A are finite, and we consider a finite horizon. Furthermore, we know that because θ is finitely randomizing, there

are only finitely many u that can be chosen by the deterministic policies at each history action pair $h^{\mathfrak{n}}, a$ for which $\theta(h^{\mathfrak{n}}, a)(u)$ gives a non-zero probability. Due to the $\mathsf{rel}^{\mathfrak{n}}(\theta)$-equivalence classes, we only have one deterministic policy per unique choice combination for all relevant histories. Combining this with the fact that there are a finite number of choices that give a non-zero probability, we can conclude that there is a finite number of deterministic policies that give a non-zero probability.

Equation (3) follows from the fact that, by construction, there is exactly one deterministic policy with a non-zero probability for each choice combination of choices available in the stochastic policies over all relevant nature histories of that stochastic policy. Each of these deterministic policies is assigned the product of the probabilities assigned to the same choices by the stochastic policy for the relevant nature histories. Summing over the deterministic policies will hence equal summing over the product of the probabilities assigned to the choices by the stochastic policy for the relevant nature histories. Since the stochastic policy assigns a probability distribution over its choices for each relevant nature history, we also get that:

$$\sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) = 1.$$

□

The following lemma is the last but vital ingredient towards the proof of Theorem 5.

Lemma 8. *$g(\theta)$ equivalent to θ:*

$$\forall \theta \in \Theta, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta} = \mu^{\pi,g(\theta)}.$$

Proof. Take arbitrary $\theta \in \Theta$. We show $\forall \pi \in \Pi \cup \Pi^{mix}, \forall \tau \in \text{Paths}^M. \mu^{\pi,\theta}(\tau) = \mu^{\pi,g(\theta)}(\tau)$ by induction on the length of the path τ. We write the length of τ as $|\tau|$.

Assume $|\tau| = 0$. Then $\tau = \langle s_I \rangle$. Then we have for $\pi \in \Pi$:

$$\begin{aligned}
\mu^{\pi,\theta}(\langle s_I \rangle) &= \eta^{\pi,\theta}(\langle s_I \rangle) \\
&= 1 \\
&= \sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot 1 \\
&= \sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot \eta^{\pi,\theta^{det}}(\langle s_I \rangle) \\
&= \mu^{\pi,g(\theta)}(\langle s_I \rangle).
\end{aligned}$$

And for $\pi^{mix} \in \Pi^{mix}$:

$$\begin{aligned}
\mu^{\pi^{mix},\theta}(\langle s_I \rangle) &= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \eta^{\pi,\theta}(\langle s_I \rangle) \\
&= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot 1 \\
&= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot 1 \\
&= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot \eta^{\pi,\theta^{det}}(\langle s_I \rangle) \\
&= \mu^{\pi^{mix},g(\theta)}(\langle s_I \rangle).
\end{aligned}$$

So for paths τ of length 0, we know that $\forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta}(\tau) = \mu^{\pi,g(\theta)}(\tau)$.

Now assume we know, given $q \in \mathbb{N}, q \geq 1$, that:

$$\forall \tau \in \text{Paths}^M. |\tau| = q - 1 \implies \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta}(\tau) = \mu^{\pi,g(\theta)}(\tau).$$

Take arbitrary $\tau \in \text{Paths}^M$ with horizon length $|\tau| = q$. Then we have:

$$\tau = \tau_{0:q-1} \oplus \langle a_{q-1}, u_{q-1}, s_q \rangle = \tau_{0:q-2} \oplus \langle a_{q-2}, u_{q-2}, s_{q-1}, a_{q-1}, u_{q-1}, s_q \rangle.$$

Then $\tau_{0:q-1} \in \text{Paths}^M$ and $|\tau_{0:q-1}| = q - 1$. By assumption, we get that:

$$\forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta}(\tau_{0:q-1}) = \mu^{\pi,g(\theta)}(\tau_{0:q-1}).$$

We need to distinguish two cases for the proof: $\pi \in \Pi$ and $\pi \in \Pi^{mix}$. We write out the more complicated case of mixed agent policies $\pi \in \Pi^{mix}$. The proof for stochastic agent policies follows along the same lines. We highlight subtle changes in the equations using either blue or red text.

$$\mu^{\pi^{mix},\theta}(\tau) = \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \eta^{\pi^{det},\theta}(\tau).$$

Unfolding $\eta^{\pi^{det},\theta}(\tau)$:

$$= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \eta^{\pi^{det},\theta}(\tau_{0:q-1}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) \cdot \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q).$$

Reordering:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \eta^{\pi^{det},\theta}(\tau_{0:q-1}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}).$$

Using the definition of μ for deterministic or stochastic agent and nature policies:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \mu^{\pi^{det},\theta}(\tau_{0:q-1}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}).$$

Using our assumption $\forall \pi \in \Pi \cup \Pi^{mix}.\ \mu^{\pi,\theta}(\tau_{0:q-1}) = \mu^{\pi,g(\theta)}(\tau_{0:q-1})$, we get:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \mu^{\pi^{det},g(\theta)}(\tau_{0:q-1}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}).$$

Unfolding the definition of μ for deterministic or stochastic agent policies and mixed nature policies:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot \eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}).$$

Multiplying by a term equal to 1:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot \eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) \cdot \frac{\theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}.$$

Using that $g(\theta)$ is a probability distribution over the set of deterministic policies, we get:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot \eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) \cdot \frac{\theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}.$$

Using that $\Theta^{det,\sim_{\mathsf{rel}^{\mathfrak{n}}(\theta)}}$ contains exactly one deterministic policy for each choice combination of available choices for θ for each history relevant for θ, we get:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot \eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) \cdot \frac{\sum_{\theta^{det} \in \Theta^{det,\sim_{\mathsf{rel}^{\mathfrak{n}}(\theta)}}} \prod_{h^{\mathfrak{n}} \in \mathsf{rel}^{\mathfrak{n}}(\theta), a \in A} \theta(h^{\mathfrak{n}}, a)(\theta^{det}(h^{\mathfrak{n}}, a)) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot \theta(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}.$$

Using the definition of g:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \sum_{\theta^{det} \in \Theta^{det}} g(\theta)(\theta^{det}) \cdot \eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1}) \cdot$$

$$\theta(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})\cdot\frac{\sum_{\theta^{det}\in\Theta^{det,\sim_{\mathsf{rel}^{\mathfrak{n}}(\theta)}}}g(\theta)(\theta^{det})\cdot\theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})}{\sum_{\theta^{det}\in\Theta^{det}}g(\theta)(\theta^{det})\cdot\theta(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})}.$$

Let $\theta^{det'}$ be an arbitrary deterministic policy in the set of relevant deterministic policies $\Theta^{det,O^{\mathfrak{n},M}(\tau_{0:q-1})}$, we get:

$$=\boldsymbol{T}(u_{q-1})(s_{q-1},a_{q-1},s_q)\cdot\sum_{\pi^{det}\in\Pi^{det}}\pi^{mix}(\pi^{det})\cdot\pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1})\cdot\sum_{\theta^{det}\in\Theta^{det}}g(\theta)(\theta^{det})\cdot\eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1})\cdot$$
$$\theta(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})\cdot\frac{\eta^{\pi^{det},\theta^{det'}}(\tau_{0:q-1})\cdot\sum_{\theta^{det}\in\Theta^{det,\sim_{\mathsf{rel}^{\mathfrak{n}}(\theta)}}}g(\theta)(\theta^{det})\cdot\theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})}{\eta^{\pi^{det},\theta^{det'}}(\tau_{0:q-1})\cdot\sum_{\theta^{det}\in\Theta^{det}}g(\theta)(\theta^{det})\cdot\theta(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})}.$$

Using Lemmas 5 and 6 and the fact that $g(\theta)(\theta^{det})=0$ when $\theta^{det}\notin\Theta^{det,\sim_{\mathsf{rel}^{\mathfrak{n}}(\theta)}}$, we get that:

$$=\boldsymbol{T}(u_{q-1})(s_{q-1},a_{q-1},s_q)\cdot\sum_{\pi^{det}\in\Pi^{det}}\pi^{mix}(\pi^{det})\cdot\pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1})\cdot\sum_{\theta^{det}\in\Theta^{det}}g(\theta)(\theta^{det})\cdot\eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1})\cdot$$
$$\theta(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})\cdot\frac{\sum_{\theta^{det}\in\Theta^{det,\sim_{\mathsf{rel}^{\mathfrak{n}}(\theta)}}}g(\theta)(\theta^{det})\cdot\eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1})\cdot\theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})}{\sum_{\theta^{det}\in\Theta^{det}}g(\theta)(\theta^{det})\cdot\eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1})\cdot\theta(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})}.$$

The denominator of the fraction is now equal to a term it is multiplied by:

$$=\boldsymbol{T}(u_{q-1})(s_{q-1},a_{q-1},s_q)\cdot\sum_{\pi^{det}\in\Pi^{det}}\pi^{mix}(\pi^{det})\cdot\pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1})\cdot$$
$$\sum_{\theta^{det}\in\Theta^{det,\sim_{\mathsf{rel}^{\mathfrak{n}}(\theta)}}}g(\theta)(\theta^{det})\cdot\eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1})\cdot\theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1}).$$

Using that $\forall\theta^{det}\in\Theta^{det}\backslash\Theta^{det,\sim_{\mathsf{rel}^{\mathfrak{n}}(\theta)}}.\,g(\theta)(\theta^{det})=0$ by definition, we get that:

$$=\boldsymbol{T}(u_{q-1})(s_{q-1},a_{q-1},s_q)\cdot\sum_{\pi^{det}\in\Pi^{det}}\pi^{mix}(\pi^{det})\cdot\pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1})\cdot\sum_{\theta^{det}\in\Theta^{det}}g(\theta)(\theta^{det})\cdot\eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1})\cdot$$
$$\theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1}).$$

Reordering:

$$=\sum_{\pi^{det}\in\Pi^{det}}\pi^{mix}(\pi^{det})\cdot\sum_{\theta^{det}\in\Theta^{det}}g(\theta)(\theta^{det})\cdot\eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1})\cdot\pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1})\cdot$$
$$\theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}),a_{q-1})(u_{q-1})\cdot\boldsymbol{T}(u_{q-1})(s_{q-1},a_{q-1},s_q).$$

Folding $\eta^{\pi^{det},\theta^{det}}(\tau)$:

$$=\sum_{\pi^{det}\in\Pi^{det}}\pi^{mix}(\pi^{det})\cdot\sum_{\theta^{det}\in\Theta^{det}}g(\theta)(\theta^{det})\cdot\eta^{\pi^{det},\theta^{det}}(\tau).$$

Using the definition of μ for mixed agent and nature policies:

$$=\mu^{\pi^{mix},g(\theta)}(\tau).$$

So if $\forall\pi\in\Pi\cup\Pi^{mix}.\,\mu^{\pi,\theta}(\tau)=\mu^{\pi,g(\theta)}(\tau)$ holds for paths of arbitrary length $q-1\in\mathbb{N}$, $\forall\pi\in\Pi\cup\Pi^{mix}.\,\mu^{\pi,\theta}(\tau)=\mu^{\pi,g(\theta)}(\tau)$ holds for paths of length q. Hence, by induction, $\forall\tau\in\text{Paths}^M,\forall\pi\in\Pi\cup\Pi^{mix}.\,\mu^{\pi,\theta}(\tau)=\mu^{\pi,g(\theta)}(\tau)$.
As $\theta\in\Theta$ was arbitrarily chosen, we conclude that:

$$\forall\theta\in\Theta,\forall\pi\in\Pi\cup\Pi^{mix}.\,\mu^{\pi,\theta}=\mu^{\pi,g(\theta)}.$$

□

We can now prove the nature case of Theorem 5:

Theorem 5 (Existence of equivalent mixed policy)**.** *Let $\mu^{\pi,\theta}\in\Delta(\mathit{Paths}^M)$ be the probability distribution over paths in the RPOMDP resulting from executing agent policy π and nature policy θ. Then:*

$$\forall\pi\in\Pi,\exists\pi^{mix}\in\Pi^{mix},\forall\theta\in\Theta\cup\Theta^{mix}.\,\mu^{\pi,\theta}=\mu^{\pi^{mix},\theta},$$
$$\forall\theta\in\Theta,\exists\theta^{mix}\in\Theta^{mix},\forall\pi\in\Pi\cup\Pi^{mix}.\,\mu^{\pi,\theta}=\mu^{\pi,\theta^{mix}}.$$

Proof. Take arbitrary $\theta \in \Theta$. By Lemma 7, we know that:

$$g(\theta) \in \Theta^{mix}.$$

Furthermore, by Lemma 8, we know that:

$$\forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta} = \mu^{\pi,g(\theta)}.$$

Hence, we know that:

$$\exists \theta^{mix} \in \Theta^{mix}, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta} = \mu^{\pi,\theta^{mix}}.$$

As $\theta \in \Theta$ was arbitrarily chosen, we conclude that:

$$\forall \theta \in \Theta, \exists \theta^{mix} \in \Theta^{mix}, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta} = \mu^{\pi,\theta^{mix}}.$$

□

Proof of Theorem 6

We define a function $f \colon \Theta^{mix} \to \Theta$ with which we will construct our equivalent stochastic policy. This function follows the construction used in [Kuhn, 1953]. Note that we cannot directly apply Kuhn's theorem, as our game is not finite. If a nature history is relevant for the mixed policy, the probability for each choice in the resulting stochastic policy will only take the probability of deterministic policies that can reach that nature history into account. Nature histories that are not relevant for the mixed policy will also not be relevant for the resulting policy, so for those histories, we can just look at the probability of all deterministic policies.

$$f(\theta^{mix})(h^{\mathtt{n}}, a)(u) = \begin{cases} \frac{\sum_{\theta^{det} \in \Theta^{det,h^{\mathtt{n}}}} \theta^{det}(h^{\mathtt{n}}, a)(u) \cdot \theta^{mix}(\theta^{det})}{\sum_{\theta^{det} \in \Theta^{det,h^{\mathtt{n}}}} \theta^{mix}(\theta^{det})} & \text{if } h^{\mathtt{n}} \in \mathsf{rel}^{\mathtt{n}}(\theta^{mix}), \\ \sum_{\theta^{det} \in \Theta^{det}} \theta^{det}(h^{\mathtt{n}}, a)(u) \cdot \theta^{mix}(\theta^{det}) & \text{if } h^{\mathtt{n}} \notin \mathsf{rel}^{\mathtt{n}}(\theta^{mix}). \end{cases}$$

where

$$\theta^{det}(h^{\mathtt{n}}, a)(u) = \begin{cases} 1 & \text{if } \theta^{det}(h^{\mathtt{n}}, a) = u, \\ 0 & \text{otherwise.} \end{cases}$$

We first show that f correctly maps to a stochastic policy (Lemma 9) and then that this resulting policy results in the same distribution over paths in the RPOMDP given any agent policy (Lemma 10).

Lemma 9 ($f(\theta^{mix})$ is a stochastic policy)**.**

$$\forall \theta^{mix} \in \Theta^{mix}. f(\theta^{mix}) \in \Theta.$$

Proof. Take arbitrary $\theta^{mix} \in \Theta^{mix}$. By construction, we have that $f(\theta^{mix}) \in H^{\mathtt{n}} \times A \to \Delta(\boldsymbol{U})$. Now to show that $f(\theta^{mix}) \in \Theta$, we must show two things: $f(\theta^{mix})$ assigns a non-zero probability to a finite number of variable assignments (finitely randomizing) and $f(\theta^{mix})$ is valid, meaning it adheres to the stickiness restrictions (see Section 3.1).

$$\forall h^{\mathtt{n}} \in H^{\mathtt{n}}, \forall a \in A. f(\theta^{mix})(h^{\mathtt{n}}, a) \text{ is finitely randomizing,} \tag{4}$$

$$\forall h^{\mathtt{n}} \in H^{\mathtt{n}}, \forall a \in A, \forall u \in f(\theta^{mix})(h^{\mathtt{n}}, a). u \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(h^{\mathtt{n}})). \tag{5}$$

Equation (4) follows from θ^{mix} being finitely randomizing by definition. Every choice in $f(\theta^{mix})$ assigns a non-zero probability to a number of variable assignments less than or equal to the number of the deterministic policies with a non-zero probability in the mixed policy. As this number of deterministic policies is finite, the resulting policy $f(\theta^{mix})$ is finitely randomizing.

Equation (5) follows from the fact that the deterministic policies used to construct $f(\theta^{mix})$ are valid policies by definition, so we have

$$\forall \theta^{det} \in \Theta^{det}, \forall h^{\mathtt{n}} \in H^{\mathtt{n}}, \forall a \in A, \theta^{det}(h^{\mathtt{n}}, a) \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(h^{\mathtt{n}})).$$

Combining this with the fact that

$$\forall h^{\mathtt{n}} \in H^{\mathtt{n}}, \forall a \in A, \forall u \in f(\theta^{mix})(h^{\mathtt{n}}, a), \exists \theta^{det} \in \Theta^{det}. \theta^{mix}(\theta^{det}) > 0 \wedge \theta^{det}(h^{\mathtt{n}}, a) = u,$$

gives us the desired result:

$$\forall h^{\mathtt{n}} \in H^{\mathtt{n}}, \forall a \in A, \forall u \in f(\theta^{mix})(h^{\mathtt{n}}, a). u \in \boldsymbol{U}^{\mathcal{P}}(\mathsf{fix}(h^{\mathtt{n}})).$$

□

Lemma 10. *$f(\theta^{mix})$ equivalent to θ^{mix}:*

$$\forall \theta^{mix} \in \Theta^{mix}, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi,\theta^{mix}} = \mu^{\pi,f(\theta^{mix})}.$$

Proof. Take arbitrary $\theta^{mix} \in \Theta^{mix}$. We show $\forall \pi \in \Pi \cup \Pi^{mix}, \forall \tau \in \text{Paths}^M . \mu^{\pi,\theta^{mix}}(\tau) = \mu^{\pi,f(\theta^{mix})}(\tau)$ by induction on the length of the path τ. We write the length of τ as $|\tau|$.

Assume $|\tau| = 0$. Then $\tau = \langle s_I \rangle$. Then we have for $\pi \in \Pi$:

$$\begin{aligned}
\mu^{\pi,\theta^{mix}}(\langle s_I \rangle) &= \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi,\theta^{det}}(\langle s_I \rangle) \\
&= \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot 1 \\
&= 1 \\
&= \eta^{\pi,f(\theta^{mix})}(\langle s_I \rangle) \\
&= \mu^{\pi,f(\theta^{mix})}(\langle s_I \rangle),
\end{aligned}$$

and for $\pi^{mix} \in \Pi^{mix}$:

$$\begin{aligned}
\mu^{\pi^{mix},\theta^{mix}}(\langle s_I \rangle) &= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi,\theta^{det}}(\langle s_I \rangle) \\
&= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot 1 \\
&= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot 1 \\
&= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \eta^{\pi,f(\theta^{mix})}(\langle s_I \rangle) \\
&= \mu^{\pi^{mix},f(\theta^{mix})}(\langle s_I \rangle).
\end{aligned}$$

So for paths τ of length 0, we know that $\forall \pi \in \Pi \cup \Pi^{mix} . \mu^{\pi,\theta^{mix}}(\tau) = \mu^{\pi,f(\theta^{mix})}(\tau)$.

Now assume we know, given $q \in \mathbb{N}, q \geq 1$, that:

$$\forall \tau \in \text{Paths}^M . |\tau| = q - 1 \implies \forall \pi \in \Pi \cup \Pi^{mix} . \mu^{\pi,\theta^{mix}}(\tau) = \mu^{\pi,f(\theta^{mix})}(\tau).$$

Take arbitrary $\tau \in \text{Paths}^M$ with horizon length $|\tau| = q$. Then we have:

$$\tau = \tau_{0:q-1} \oplus \langle a_{q-1}, u_{q-1}, s_q \rangle = \tau_{0:q-2} \oplus \langle a_{q-2}, u_{q-2}, s_{q-1}, a_{q-1}, u_{q-1}, s_q \rangle.$$

Then $\tau_{0:q-1} \in \text{Paths}^M$ and $|\tau_{0:q-1}| = q - 1$. By assumption, we get that:

$$\forall \pi \in \Pi \cup \Pi^{mix} . \mu^{\pi,\theta^{mix}}(\tau_{0:q-1}) = \mu^{\pi,f(\theta^{mix})}(\tau_{0:q-1}).$$

We also assume that $O^{\mathfrak{n},M}(\tau_{0:q-1}) \in \mathsf{rel}^{\mathfrak{n}}(\theta^{mix})$. If $O^{\mathfrak{n},M}(\tau_{0:q-1}) \notin \mathsf{rel}^{\mathfrak{n}}(\theta^{mix})$, so if the history of the path's prefix is not relevant for the mixed policy, the path will not be generated by the mixed policy or the stochastic policies. Then we trivially have:

$$\forall \pi \in \Pi \cup \Pi^{mix} . \mu^{\pi,\theta^{mix}}(\tau) = 0 = \mu^{\pi,f(\theta^{mix})}(\tau).$$

We need to distinguish two cases for the proof: $\pi \in \Pi$ and $\pi \in \Pi^{mix}$. We write out the more complicated case: $\pi \in \Pi^{mix}$. The other proof follows along the same lines. We indicate hard-to-read changes in the equations using either blue or red text.

$$\mu^{\pi^{mix},\theta^{mix}}(\tau) = \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det},\theta^{det}}(\tau).$$

Unfolding $\eta^{\pi^{det},\theta^{det}}(\tau)$:

$$\begin{aligned}
&= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det},\theta^{det}}(\tau_{0:q-1}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \\
&\quad \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) \cdot \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q).
\end{aligned}$$

Reordering:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot$$

$$\sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1}) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}).$$

Multiplying by a term equal to 1:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot$$

$$\sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1}) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) \cdot \frac{\sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1})}{\sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1})}.$$

Reordering:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1}) \cdot$$

$$\frac{\sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1}) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1})}.$$

Using the definition of μ for deterministic or stochastic agent policies and mixed nature policies:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \mu^{\pi^{det}, \theta^{mix}}(\tau_{0:q-1}) \cdot$$

$$\frac{\sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1}) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1})}.$$

Using lemma 5, we get:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \mu^{\pi^{det}, \theta^{mix}}(\tau_{0:q-1}) \cdot$$

$$\frac{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n},M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1}) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n},M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det}}(\tau_{0:q-1})}.$$

Let $\theta^{det'}$ be an arbitrary deterministic policy in the set of relevant deterministic policies $\Theta^{det, O^{\mathfrak{n},M}(\tau_{0:q-1})}$. Using lemma 6, we get:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \mu^{\pi^{det}, \theta^{mix}}(\tau_{0:q-1}) \cdot$$

$$\frac{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n},M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det'}}(\tau_{0:q-1}) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n},M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det}) \cdot \eta^{\pi^{det}, \theta^{det'}}(\tau_{0:q-1})}.$$

Reordering:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \mu^{\pi^{det}, \theta^{mix}}(\tau_{0:q-1}) \cdot$$

$$\frac{\eta^{\pi^{det}, \theta^{det'}}(\tau_{0:q-1}) \cdot \sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n},M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det}) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\eta^{\pi^{det}, \theta^{det'}}(\tau_{0:q-1}) \cdot \sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n},M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det})}.$$

Now we can simplify the fraction:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \mu^{\pi^{det}, \theta^{mix}}(\tau_{0:q-1}) \cdot$$

$$\frac{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n},M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det}) \cdot \theta^{det}(O^{\mathfrak{n},M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n},M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det})}.$$

Using our assumption $\forall \pi \in \Pi \cup \Pi^{mix} \mu^{\pi, \theta^{mix}}(\tau_{0:q-1}) = \mu^{\pi, f(\theta^{mix})}(\tau_{0:q-1})$, we get:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a},M}(\tau_{0:q-1}))(a_{q-1}) \cdot \mu^{\pi^{det}, f(\theta^{mix})}(\tau_{0:q-1}) \cdot$$

$$\frac{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n}, M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det}) \cdot \theta^{det}(O^{\mathfrak{n}, M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n}, M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det})}.$$

Using the definition of μ for deterministic or stochastic agent and nature policies:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a}, M}(\tau_{0:q-1}))(a_{q-1}) \cdot \eta^{\pi^{det}, f(\theta^{mix})}(\tau_{0:q-1}) \cdot$$
$$\frac{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n}, M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det}) \cdot \theta^{det}(O^{\mathfrak{n}, M}(\tau_{0:q-1}), a_{q-1})(u_{q-1})}{\sum_{\theta^{det} \in \Theta^{det, O^{\mathfrak{n}, M}(\tau_{0:q-1})}} \theta^{mix}(\theta^{det})}.$$

Using the definition of f and our assumption the $O^{\mathfrak{n}, M}(\tau_{0:q-1}) \in \mathsf{rel}^{\mathfrak{n}}(\theta^{mix})$, we get:

$$= \boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q) \cdot \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \pi^{det}(O^{\mathfrak{a}, M}(\tau_{0:q-1}))(a_{q-1}) \cdot \eta^{\pi^{det}, f(\theta^{mix})}(\tau_{0:q-1}) \cdot$$
$$f(\theta^{mix})(O^{\mathfrak{n}, M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}).$$

Reordering:

$$= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \eta^{\pi^{det}, f(\theta^{mix})}(\tau_{0:q-1}) \cdot \pi^{det}(O^{\mathfrak{a}, M}(\tau_{0:q-1}))(a_{q-1}) \cdot f(\theta^{mix})(O^{\mathfrak{n}, M}(\tau_{0:q-1}), a_{q-1})(u_{q-1}) \cdot$$
$$\boldsymbol{T}(u_{q-1})(s_{q-1}, a_{q-1}, s_q).$$

Folding $\eta^{\pi^{det}, f(\theta^{mix})}(\tau)$:

$$= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \eta^{\pi^{det}, f(\theta^{mix})}(\tau).$$

Using the definition of μ for mixed agent policies and deterministic or stochastic nature policies:

$$= \mu^{\pi^{mix}, f(\theta^{mix})}(\tau).$$

So if $\forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi, \theta^{mix}}(\tau) = \mu^{\pi, f(\theta^{mix})}(\tau)$ holds for paths of arbitrary length $q-1 \in \mathbb{N}$, $\forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi, \theta^{mix}}(\tau) = \mu^{\pi, f(\theta^{mix})}(\tau)$ holds for paths of length q. Hence, by induction, $\forall \tau \in \text{Paths}^M, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi, \theta^{mix}}(\tau) = \mu^{\pi, f(\theta^{mix})}(\tau)$. As $\theta^{mix} \in \Theta^{mix}$ was arbitrarily chosen, we conclude that:

$$\forall \theta^{mix} \in \Theta^{mix}, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi, \theta^{mix}} = \mu^{\pi, f(\theta^{mix})}.$$

□

We can now prove the nature case of Theorem 6:

Theorem 6 (Existence of equivalent stochastic policy)**.** *Let $\mu^{\pi, \theta} \in \Delta(\mathit{Paths}^M)$ be the probability distribution over paths in the RPOMDP resulting from executing agent policy π and nature policy θ. Then:*

$$\forall \pi^{mix} \in \Pi^{mix}, \exists \pi \in \Pi, \forall \theta \in \Theta \cup \Theta^{mix}. \mu^{\pi^{mix}, \theta} = \mu^{\pi, \theta},$$
$$\forall \theta^{mix} \in \Theta^{mix}, \exists \theta \in \Theta, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi, \theta^{mix}} = \mu^{\pi, \theta}.$$

Proof. Take arbitrary $\theta^{mix} \in \Theta^{mix}$. By Lemma 9, we know that:

$$f(\theta^{mix}) \in \Theta.$$

Furthermore, by Lemma 10, we know that:

$$\forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi, \theta^{mix}} = \mu^{\pi, f(\theta^{mix})}.$$

Hence, we know that:

$$\exists \theta \in \Theta, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi, \theta^{mix}} = \mu^{\pi, \theta}.$$

As $\theta^{mix} \in \Theta^{mix}$ was arbitrarily chosen, we conclude that:

$$\forall \theta^{mix} \in \Theta^{mix}, \exists \theta \in \Theta, \forall \pi \in \Pi \cup \Pi^{mix}. \mu^{\pi, \theta^{mix}} = \mu^{\pi, \theta}.$$

□

G.4 Convex Semi-Infinite Game

Although we follow the occupancy game construction from [Delage *et al.*, 2023], their proof for Nash equilibrium existence requires both players to have convex subsets of a Euclidean space as their policy space. Our set of nature policies does not meet this requirement, as the number of deterministic nature policies is infinite, and therefore, the mixed policy space is infinite-dimensional. Instead, we show that our occupancy game is a convex semi-infinite game as defined in [Lopez and Vercher, 1986] and given below:

Definition 22 (Convex semi-infinite game [Lopez and Vercher, 1986])**.** *Given an arbitrary, possibly infinite set T, a convex semi-infinite game is a zero-sum game with policy sets Γ and C of the following types:*

- $\Gamma = \{\vec{\lambda} = (\lambda_t)_{t\in T} \mid$ *only finitely many* $\lambda_t \neq 0, \lambda_t \geq 0,$ *and* $\sum_{t\in T} \lambda_t = 1\}$.
- $C =$ *a nonempty closed convex set in* $\mathbb{R}^n$ *with* $n \in \mathbb{N}$ *a finite number.*

And a family of convex functions with finite values $\mathcal{F}_T = \{F_t \colon \mathbb{R}^n \to \mathbb{R} \mid t \in T\}$, such that the value function of the convex semi-infinite game $W \colon C \times \Gamma \to \mathbb{R}$ is:

$$W(\vec{x}, \vec{\lambda}) = \sum_{t\in T} \lambda_t F_t(\vec{x}), \text{ with } \vec{\lambda} \in \Gamma \text{ and } \vec{x} \in C.$$

Theorem 7 (Occupancy game is convex semi-infinite)**.** *Given an RPOMDP M and horizon $K \in \mathbb{N}$, the corresponding occupancy game OG (Definition 19) is an convex semi-infinite game, where:*

- *T is the set of deterministic nature policies $\Theta^{det}_{0:K-1}$.*
- *Γ is the set of mixed nature policies $\Theta^{mix}_{0:K-1}$.*
- *C is the set of mixed agent policies $\Pi^{mix}_{0:K-1}$.*
- *$F_{\theta^{det}}(\vec{x}) = \sum_{\pi^{det}\in\Pi^{det}_{0:K-1}} x_{\pi^{det}} \cdot V^{\pi^{det},\theta^{det}}$ with $\theta^{det} \in \Theta^{det}_{0:K-1} = T$ and $\vec{x} \in \mathbb{R}^n$ where n is the finite number of deterministic agent policies $|\Pi^{det}_{0:K-1}|$.*

Note that we omitted the history length indication for the policies for readability purposes. The history lengths on which the policies are defined can be derived from the policy sets from which they are taken. See Table 2 in Appendix A for the notation glossary.

We prove Theorem 7 by proving several smaller lemmas, showing that the suggested mapping of the occupancy game to the convex semi-infinite game definition is correct. Lemma 11 shows that the set of mixed agent policies $\Pi^{mix}_{0:K-1}$ meets the conditions for policy set C of the convex semi-infinite game definition.

Lemma 11. *$\Pi^{mix}_{0:K-1}$ is a nonempty closed convex set in $\mathbb{R}^n$ with $n = |\Pi^{det}_{0:K-1}|$ a finite number.*

Proof. By definition, $\Pi^{mix}_{0:K-1} = \Delta(\Pi^{det}_{0:K-1})$. Since the set of actions A, the set of agent observation $Z^{\mathfrak{a}}_{\bullet}$, and the set of public observation $Z_\circ$ are all finite and nonempty, the number of deterministic policies $|\Pi^{det}_{0:K-1}|$ is finite and nonempty and less than or equal to

$$\sum_{t=0}^{K} (|A| \cdot |Z^{\mathfrak{a}}_{\bullet}| \cdot |Z_\circ|)^t \cdot |A|.$$

Let n be the actual number of different deterministic policies in our RPOMDP. Then we know the set of mixed policies $\Pi^{mix}_{0:K-1} \subseteq \mathbb{R}^n$ with n a finite number. Finally, since the set of mixed policies is the probability simplex $\Delta(\Pi^{det}_{0:K-1}) = \Delta(\mathbb{R}^n)$, we know that it is a closed and convex set. □

As no restrictions are given on T, we can take the infinite set of deterministic nature policies $\Theta^{det}_{0:K-1}$. Lemma 12 shows that the mixed nature policies then meet the conditions for policy set Γ of the convex semi-infinite game definition.

Lemma 12. *$\Theta^{mix}_{0:K-1}$ is the set of finite probability distributions over the set of deterministic nature policies $\Theta^{det}_{0:K-1}$:*

$$\Theta^{mix}_{0:K-1} = \{\vec{\lambda} = (\lambda_{\theta^{det}})_{\theta^{det}\in\Theta^{det}_{0:K-1}} \mid \text{only finitely many } \lambda_{\theta^{det}} \neq 0, \lambda_{\theta^{det}} \geq 0, \text{ and } \sum_{\theta^{det}\in\Theta^{det}_{0:K-1}} \lambda_{\theta^{det}} = 1\}.$$

Proof. By definition, the set of mixed nature policies $\Theta^{mix}_{0:K-1}$ is the set of probability distributions over the set of deterministic nature policies $\Theta^{det}_{0:K-1}$. As stated in Section 2, we only consider finite probability distributions over infinite sets. The set of mixed nature policies is hence also restricted to the finite probability distributions. This means $\vec{\lambda} = \theta^{mix} \in \Theta^{mix}_{0:K-1}$ with $\lambda_{\theta^{det}} = \theta^{mix}(\theta^{det})$. □

Recall the family of functions $F_{\theta^{det}} : \mathbb{R}^n \to \mathbb{R}$ defined on the set of deterministic nature policies $\Theta^{det}_{0:K-1}$ in Theorem 7:

$$F_{\theta^{det}}(x_{\pi^{det}}) = \sum_{\pi^{det} \in \Pi^{det}_{0:K-1}} x_{\pi^{det}} \cdot V^{\pi^{det},\theta^{det}}.$$

The next two lemmas show that this family of functions $F_{\theta^{det}}$ is convex (Lemma 13) and all functions in the family have a finite value (Lemma 14).

Lemma 13. $\forall \theta^{det} \in \Theta^{det}_{0:K-1}$. *$F_{\theta^{det}}$ is a convex function.*

Proof. Take arbitrary $\theta^{det} \in \Theta^{det}_{0:K-1}$, $\vec{x}, \vec{y} \in \mathbb{R}^n$, and $\alpha \in [0, 1]$ then:

$$\begin{aligned}
F_{\theta^{det}}(\alpha \cdot \vec{x} + (1-\alpha) \cdot \vec{y}) &= \sum_{\pi^{det} \in \Pi^{det}} (\alpha \cdot \vec{x} + (1-\alpha) \cdot \vec{y})_{\pi^{det}} \cdot V^{\pi^{det},\theta^{det}} \\
&= \sum_{\pi^{det} \in \Pi^{det}} (\alpha \cdot x_{\pi^{det}} + (1-\alpha) \cdot y_{\pi^{det}}) \cdot V^{\pi^{det},\theta^{det}} \\
&= \sum_{\pi^{det} \in \Pi^{det}} \alpha \cdot x_{\pi^{det}} \cdot V^{\pi^{det},\theta^{det}} + (1-\alpha) \cdot y_{\pi^{det}} \cdot V^{\pi^{det},\theta^{det}} \\
&= \sum_{\pi^{det} \in \Pi^{det}} \left\{\alpha \cdot x_{\pi^{det}} \cdot V^{\pi^{det},\theta^{det}}\right\} + \sum_{\pi^{det} \in \Pi^{det}} \left\{(1-\alpha) \cdot y_{\pi^{det}} \cdot V^{\pi^{det},\theta^{det}}\right\} \\
&= \alpha \cdot \sum_{\pi^{det} \in \Pi^{det}} \left\{x_{\pi^{det}} \cdot V^{\pi^{det},\theta^{det}}\right\} + (1-\alpha) \cdot \sum_{\pi^{det} \in \Pi^{det}} \left\{y_{\pi^{det}} \cdot V^{\pi^{det},\theta^{det}}\right\} \\
&= \alpha \cdot F_{\theta^{det}}(\vec{x}) + (1-\alpha) \cdot F_{\theta^{det}}(\vec{y}).
\end{aligned}$$

So $\forall_{\theta^{det} \in \Theta^{det}}$. $F_{\theta^{det}}$ is a convex function. □

Lemma 14. *Given $\theta^{det} \in \Theta^{det}$, $\vec{x} \in \mathbb{R}^n$:*

$$F_{\theta^{det}}(\vec{x}) \textit{ has a finite value.}$$

Proof. $\forall \theta^{det} \in \Theta^{det}, \forall \pi^{det} \in \Pi^{det}$ the value of $V^{\pi^{det}_i,\theta^{det}}$ is finite, as it is bounded by $K \cdot (\max_{s,a \in S \times A} R(s,a))$. Furthermore, $\vec{x} \in \mathbb{R}^n$ can only take finite values. Together, this shows that:

$$\forall \theta^{det} \in \Theta^{det}, \forall \pi^{det} \in \Pi^{det}. F_{\theta^{det}}(\vec{x}) \text{ has a finite value.}$$

□

Finally, Lemma 15 shows that the value function $W : \Pi^{mix}_{0:K-1} \times \Theta^{mix}_{0:K-1} \to \mathbb{R}$ of the convex semi-infinite game constructed as in Theorem 7 is equivalent to the value function $V : \Pi^{mix}_{0:K-1} \times \Theta^{mix}_{0:K-1} \to \mathbb{R}$ of the occupancy game.

Lemma 15. *The value function of the convex semi-infinite game is equivalent to the value function of the occupancy game.*

$$\forall \pi^{mix} \in \Pi^{mix}_{0:K-1}, \forall \theta^{mix} \in \Theta^{mix}_{0:K-1}. W(\pi^{mix}, \theta^{mix}) = V^{\pi^{mix},\theta^{mix}}.$$

Proof. Recall the definition of the value function of a convex semi-infinite game:

$$W(\pi^{mix}, \theta^{mix}) = \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot F_{\theta^{det}}(\pi^{mix}).$$

By construction, the value function of our occupancy game is the same as that of our original RPOMDP. However, as shown in Appendix G.3, we can reason with mixed policies, giving us the following value function:

$$V^{\pi^{mix},\theta^{mix}} = \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot V^{\pi^{det},\theta^{det}}.$$

Take arbitrary mixed agent and nature policies $\pi^{mix} \in \Pi^{mix}_{0:K-1}, \theta^{mix} \in \Theta^{mix}_{0:K-1}$. Then:

$$\begin{aligned}
W(\pi^{mix}, \theta^{mix}) &= \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot F_{\theta^{det}}(\pi^{mix}) \\
&= \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \pi^{mix}(\pi^{det}) \cdot V^{\pi^{det},\theta^{det}}
\end{aligned}$$

$$= \sum_{\pi^{det} \in \Pi^{det}} \pi^{mix}(\pi^{det}) \cdot \sum_{\theta^{det} \in \Theta^{det}} \theta^{mix}(\theta^{det}) \cdot V^{\pi^{det}, \theta^{det}}$$
$$= V^{\pi^{mix}, \theta^{mix}}.$$

□

Lemma 15 is the final step in proving Theorem 7. We conclude that our occupancy game is a convex semi-infinite game.

The final step in proving the existence of a finite horizon Nash equilibrium in our RPOMDPs follows from [Lopez and Vercher, 1986, Theorem 3.2], stating that in a convex semi-infinite game, if the convex functions and the convex agent policy set have no common direction of recession, then a Nash equilibrium and an optimal strategy for the agent exist.

Lemma 16. *In the convex semi-infinite game of our occupancy game, the convex functions $F_{\theta^{det}}$ with $\theta^{det} \in \Theta^{det}$ and the set of mixed agent policies Π^{mix} have no common direction of recession.*

Proof. As our set of agent policies is a convex polytope in $\mathbb{R}^n$, it is a closed and bounded convex subset of $\mathbb{R}^n$. Then, the recession cone consists only of the zero vector [Zalinescu, 2002]. The zero vector is also trivially contained in the recession cones of the convex functions. Therefore, we know that the convex functions and the convex agent policy set have no common direction of recession. □

As shown in Lemma 16, our occupancy game meets the condition given in [Lopez and Vercher, 1986]. It hence follows that a Nash equilibrium and an optimal strategy for the agent exist and that the saddle point condition holds, proving Theorem 3.

G.5 Nature First

When reasoning with the nature first semantics, the nature policy no longer relies on the last action of the agent. This influences the proof of the sufficient statistic as follows:

$$\mathrm{OS}_{\{\pi,\theta\}_{0:t}}(\langle h_t, a_t, u_t, z^{\mathtt{a}}_{\bullet}, z^{\mathtt{n}}_{\bullet}, z_{\circ}\rangle) \stackrel{\text{def}}{=} \langle \sigma_{\{\pi,\theta\}_{0:t}}(\langle h_t, a_t, u_t, z^{\mathtt{a}}_{\bullet}, z^{\mathtt{n}}_{\bullet}, z_{\circ}\rangle), \theta_{0:t}\rangle,$$

where:

$$\theta_{0:t} \stackrel{\text{def}}{=} \langle \theta_{0:t-1}, \theta_t \rangle.$$
$$\sigma_{\{\pi,\theta\}_{0:t}}(\langle h_t, a_t, u_t, z^{\mathtt{a}}_{\bullet}, z^{\mathtt{n}}_{\bullet}, z_{\circ}\rangle) \stackrel{\text{def}}{=} \Pr(h_t, a_t, u_t, z^{\mathtt{a}}_{\bullet}, z^{\mathtt{n}}_{\bullet}, z_{\circ} \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s \in \mathcal{S}^{\mathtt{n}}} \sum_{s' \in \mathcal{S}^{\mathtt{n}}} \Pr(z^{\mathtt{a}}_{\bullet}, z^{\mathtt{n}}_{\bullet}, z_{\circ} \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(h_t, a_t, u_t, s \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s \in \mathcal{S}^{\mathtt{n}}} \sum_{s' \in \mathcal{S}^{\mathtt{n}}} \Pr(z^{\mathtt{a}}_{\bullet}, z^{\mathtt{n}}_{\bullet}, z_{\circ} \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, s, \pi_{0:t}, \theta_{0:t}) \Pr(h_t, a_t, s \mid \pi_{0:t}, \theta_{0:t}).$$

The chance of a nature action only depends on nature's policy at time t and the history:

$$= \sum_{s \in \mathcal{S}^{\mathtt{n}}} \sum_{s' \in \mathcal{S}^{\mathtt{n}}} \Pr(z^{\mathtt{a}}_{\bullet}, z^{\mathtt{n}}_{\bullet}, z_{\circ} \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, \theta_t) \Pr(h_t, a_t, s \mid \pi_{0:t}, \theta_{0:t})$$
$$= \sum_{s \in \mathcal{S}^{\mathtt{n}}} \sum_{s' \in \mathcal{S}^{\mathtt{n}}} \Pr(z^{\mathtt{a}}_{\bullet}, z^{\mathtt{n}}_{\bullet}, z_{\circ} \mid s') \Pr(s' \mid a_t, u_t, s) \Pr(u_t \mid h_t, a_t, \theta_t) \Pr(a_t \mid h_t, \pi_t) \Pr(s \mid h_t) \Pr(h_t \mid \pi_{0:t-1}, \theta_{0:t-1})$$
$$= \sum_{s \in \mathcal{S}^{\mathtt{n}}} \sum_{s' \in \mathcal{S}^{\mathtt{n}}} \mathcal{O}^{\mathtt{a}}(s', z^{\mathtt{a}}_{\bullet}, z_{\circ},) \mathcal{O}^{\mathtt{n}}(s', z^{\mathtt{n}}_{\bullet}, z_{\circ}) \mathcal{T}^{\mathtt{a}}(\mathcal{T}^{\mathtt{n}}(s, u_t), a_t, s') \theta_t(h^{\mathtt{n}}_t, u_t) \pi_t(h^{\mathtt{a}}_t, a_t) b(s, h_t) \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t).$$

We can hence still compute the successor occupancy state using only the previous occupancy state $\mathrm{OS}_{\{\pi,\theta\}_{0:t-1}} = \langle \sigma_{\{\pi,\theta\}_{0:t-1}}, \theta_{0:t-1} \rangle$ and policies π_t, θ_t at time t. The expected reward proof requires no modifications.

We define the nature first OG as follows:

Definition 23 (OG)**.** *Given a POSG as defined in Definition 18 $\langle \mathcal{S}^{\mathtt{a}}, \mathcal{S}^{\mathtt{n}}, \mathcal{A}^{\mathtt{a}}, \mathcal{A}^{\mathtt{n}}, \mathcal{T}, \mathcal{R}, \mathcal{Z}^{\mathtt{a}}, \mathcal{Z}^{\mathtt{n}}, \mathcal{O}^{\mathtt{a}}, \mathcal{O}^{\mathtt{n}} \rangle$, and a horizon $K \in \mathbb{N}$, we define the OG as a tuple $(\mathsf{S}^{\mathtt{a}}, \mathsf{S}^{\mathtt{n}}, \mathsf{A}^{\mathtt{a}}, \mathsf{A}^{\mathtt{n}}, \mathsf{T}, \mathsf{R})$ where the sets of states and actions are defined as follows: $\mathsf{S}^{\mathtt{a}} = \bigcup_{t=0}^{K-1} (\bigcup_{\pi_{0:t} \in \Pi_{0:t}} \bigcup_{\theta_{0:t} \in \Theta_{0:t}} OS_{\{\pi,\theta\}_{0:t}} \times \Theta_{t+1})$ is the infinite set of agent states, and $\mathsf{S}^{\mathtt{n}} = \bigcup_{t=0}^{K-1} \bigcup_{\pi_{0:t} \in \Pi_{0:t}} \bigcup_{\theta_{0:t} \in \Theta_{0:t}} OS_{\{\pi,\theta\}_{0:t}}$ the infinite set of nature states; $\mathsf{A}^{\mathtt{a}} = \bigcup_{t=0}^{K-1} \Pi_t$ is the infinite set of agent actions, and $\mathsf{A}^{\mathtt{n}} = \bigcup_{t=0}^{K-1} \Theta_t$ the infinite set of nature actions; The transition and reward functions are then defined as:*

- $\mathsf{T} = \mathsf{T}^{\mathfrak{a}} \cup \mathsf{T}^{\mathfrak{n}}$*, the transition function, where:*
  - $\mathsf{T}^{\mathfrak{a}} \colon \mathsf{S}^{\mathfrak{a}} \times \mathsf{A}^{\mathfrak{a}} \hookrightarrow \mathsf{S}^{\mathfrak{n}}$ *the agent's transition function.*
  - $\mathsf{T}^{\mathfrak{n}} \colon \mathsf{S}^{\mathfrak{n}} \times \mathsf{A}^{\mathfrak{n}} \hookrightarrow \mathsf{S}^{\mathfrak{a}}$ *nature's transition function.*
- $\mathsf{R} \colon \mathsf{S}^{\mathfrak{a}} \times \mathsf{A}^{\mathfrak{a}} \to \mathbb{R}$ *the reward function.*

Where:

- $\mathsf{R}(\langle \sigma_{\{\pi,\theta\}_{0:t}}, \theta_{0:t} \rangle, \pi_{t+1}) = \sum_{s \in \mathcal{S}^{\mathfrak{a}}} \sum_{a \in \mathcal{A}^{\mathfrak{a}}} \Big\{ \mathcal{R}(s,a) \cdot \sum_{h_{t+1} \in H_{t+1}(\theta_{0:t})} \big\{ \pi_{t+1}(h_{t+1}, a) b(s, h_{t+1}) \sigma_{\{\pi,\theta\}_{0:t}}(h_{t+1}) \big\} \Big\}$.
- $\mathsf{T}^{\mathfrak{n}}(\langle \sigma_{\{\pi,\theta\}_{0:t}}, \theta_{0:t} \rangle, \theta_{t+1}) = \langle \langle \sigma_{\{\pi,\theta\}_{0:t}}, \theta_{0:t} \rangle, \theta_{t+1} \rangle$.
- $\mathsf{T}^{\mathfrak{a}}(\langle \langle \sigma_{\{\pi,\theta\}_{0:t}}, \theta_{0:t} \rangle, \theta_{t+1} \rangle, \pi_{t+1}) = \langle \sigma_{\{\pi,\theta\}_{0:t+1}}, \theta_{0:t+1} \rangle$*, where:*
  - $\theta_{0:t+1} = \theta_{0:t} \oplus \theta_{t+1}$.
  - $\forall h_{t+1} \in H_{t+1}(\theta_{0:t}), \forall a \in \mathcal{A}^{\mathfrak{a}}, \forall u \in \mathcal{A}^{\mathfrak{n}}, \forall z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ} \in Z^{\mathfrak{a}}_{\bullet} \times Z^{\mathfrak{n}}_{\bullet} \times Z_{\circ}, \sigma_{\{\pi,\theta\}_{0:t+1}}(\langle h_{t+1}, a, u, z^{\mathfrak{a}}_{\bullet}, z^{\mathfrak{n}}_{\bullet}, z_{\circ} \rangle) =$

$$\sum_{s \in \mathcal{S}^{\mathfrak{n}}} \sum_{s' \in \mathcal{S}^{\mathfrak{n}}} \mathcal{O}^{\mathfrak{a}}(s', z^{\mathfrak{a}}_{\bullet}, z_{\circ},) \mathcal{O}^{\mathfrak{n}}(s', z^{\mathfrak{n}}_{\bullet}, z_{\circ}) \mathcal{T}^{\mathfrak{a}}(\mathcal{T}^{\mathfrak{n}}(s, u_t), a_t, s') \theta_t(h^{\mathfrak{n}}_t, u_t) \pi_t(h^{\mathfrak{a}}_t, a_t) b(s, h_t) \sigma_{\{\pi,\theta\}_{0:t-1}}(h_t).$$

Where $b(s, h_t)$ is the belief computed by t belief updates given the joint history h_t. Where $H_t(\theta_{0:t-1}) \subset H_t$ is the subset with $u_i \in \boldsymbol{U}$ determined by θ_i given the history $h^{\mathfrak{n}}_{0:i-1}$ and action a. This is a finite subset of the infinite set of possible joint histories.

To show that for every stochastic policy there exists a mixed policy that behaves equivalently and vice versa in the nature first setting, the proofs follow the same steps as in Appendix G.3 for the agent first policies. The only required changes are to remove the agent action input for the nature policies and the corresponding $\forall a \in A$.

The nature first OG still meets all requirements for having an optimal value, which can be shown by following the same proof steps as for the agent first OG in Appendix G.4.