Best-Effort Policies for Robust Markov Decision Processes

Alessandro Abate, Thom Badings, Giuseppe De Giacomo, Francesco Fabiano

Department of Computer Science, University of Oxford, Oxford, United Kingdom
{alessandro.abate, thom.badings, giuseppe.degiacomo, francesco.fabiano}@cs.ox.ac.uk

Abstract

We study the common generalization of Markov decision processes (MDPs) with sets of transition probabilities, known as robust MDPs (RMDPs). A standard goal in RMDPs is to compute a policy that maximizes the expected return under an adversarial choice of the transition probabilities. If the uncertainty in the probabilities is independent between the states, known as s-rectangularity, such optimal robust policies can be computed efficiently using robust value iteration. However, there might still be multiple optimal robust policies, which, while equivalent with respect to the worst-case, reflect different expected returns under non-adversarial choices of the transition probabilities. Hence, we propose a refined policy selection criterion for RMDPs, drawing inspiration from the notions of *dominance* and *best-effort* in game theory. Instead of seeking a policy that only maximizes the worst-case expected return, we additionally require the policy to achieve a *maximal* expected return under different (*i.e.*, not fully adversarial) transition probabilities. We call such a policy an *optimal robust best-effort* (ORBE) policy. We prove that ORBE policies always exist, characterize their structure, and present an algorithm to compute them with a manageable overhead compared to standard robust value iteration. ORBE policies offer a principled tie-breaker among optimal robust policies. Numerical experiments show the feasibility of our approach.

1 Introduction

Markov decision processes (MDPs) are the standard model for sequential decision making in stochastic environments and are ubiquitous in artificial intelligence (AI) (Russell and Norvig 2010), operations research (Davis 2018), control theory (Åström 2012), and robotics (Hanheide et al. 2017). Within AI, MDPs are at the core of many model-based reinforcement learning methods (Moerland et al. 2023). Solving an MDP amounts to computing a *policy* (or *strategy*) for the agent, *i.e.*, a mapping from states to actions, that maximizes a particular performance value, such as the expected (discounted) cumulative reward (Puterman 1994).

Robust MDPs. A fundamental limitation of MDPs is the requirement to specify transition probabilities precisely. In practice, accurately determining these probabilities can be challenging, especially if parameters are uncertain or if the model is learned from data (Badings et al. 2023b). Moreover, optimal policies may be sensitive to small changes in the

transition probabilities (Mannor et al. 2004). To address this issue, *robust MDPs* (RMDPs) generalize MDPs by allowing for *sets of transition probabilities* (Iyengar 2005; Nilim and Ghaoui 2005; Wiesemann, Kuhn, and Rustem 2013). That is, instead of assigning precise probabilities between 0 and 1, the transitions in an RMDP are described by a set of feasible probabilities, called the *uncertainty set* of the RMDP.

The standard objective in an RMDP is to compute an *optimal robust policy*, defined as a policy that *maximizes* the expected return under the *minimizing* (*i.e.*, worst-case) transition probabilities in the uncertainty set. Unfortunately, computing optimal robust policies under general uncertainty sets is NP-hard (Wiesemann, Kuhn, and Rustem 2013). To ensure tractability, uncertainty sets are commonly assumed to be convex as well as independent between the states and/or actions of the RMDP, referred to as *rectangularity* of the uncertainty set. Under these assumptions, optimal robust policies can be computed, *e.g.*, using robust value iteration.

The adversarial nature of RMDPs. When computing an optimal robust policy, the choice of transition probabilities is inherently adversarial. However, in many scenarios, the choice of transition probabilities is *not* actively working against the agent, making this assumption overly conservative. Take, for example, an autonomous drone flying through uncertain wind conditions. Clearly, the wind conditions do not depend on the drone's control policy, so reasoning solely about the worst-case conditions might be too conservative. Moreover, multiple optimal robust policies may exist, even though their performance under non-adversarial conditions may differ. We thus raise the vital question: can we compute a policy that is optimal in the worst case, but also "is best" when the environment does not act fully adversarially?

Best-effort policies. To address the limitations of purely adversarial reasoning in RMDPs, we draw inspiration from advances in reactive stochastic games (Aminof et al. 2023; Giacomo, Favorito, and Silo 2024). In this framework, a policy is deemed *winning*, *dominant*, or *best-effort* if it succeeds against *all*, the *maximum* subset, or a *maximal* subset of the environment policies, respectively. Yet, these papers consider games where only the graph of the model is known and the probabilities are unconstrained, as opposed to RMDPs, where the uncertainty is captured by bounded sets of distributions.

In this paper, we leverage the concepts of dominance and

best-effort to define a refined policy selection criterion for RMDPs, which we term *optimal robust best-effort* (ORBE). An ORBE policy satisfies two properties: (1) it achieves an optimal expected return under the *worst-case* transition probabilities; and (2) it is not dominated by any other policy, *i.e.*, is best-effort. Here, one policy is said to dominate another if it performs at least as well across the entire uncertainty set and strictly better in at least one instance of the transition probabilities from the uncertainty set. This best-effort perspective offers a principled tie-breaker among optimal robust policies, favoring those achieving a maximal expected return under non-adversarial transition probabilities. Thus, ORBE policies preserve robust optimality—unlike approaches that update the uncertainty set—while also improving performance in non–fully adversarial environments.

Contributions. We introduce the class of optimal robust best-effort (ORBE) policies for RMDPs. These policies combine the worst-case guarantees of standard robust policies with the refinement offered by best-effort reasoning, ensuring strong performance even when the environment is not fully adversarial. Specifically, our key contributions are as follows:

- We formalize the notions of dominant and best-effort policies within the context of RMDPs (Sect. 3).
- We present a full characterization of ORBE policies and an efficient algorithm to compute them with small overhead to standard robust value iteration (Sects. 4 and 5).
- We empirically demonstrate the feasibility of our techniques as a tie-breaker in robust value iteration (Sect. 6).

We postpone a detailed discussion of related work to Sect. 7.

2 Preliminaries

We write $\langle u, v \rangle \coloneqq \sum_{x \in X} u(x)v(x)$ for the dot product between the functions $u, v \colon X \to \mathbb{R}$. The cardinality of a set X is written as $|X|$. A probability distribution over a set X is a function $\mu \colon X \to [0, 1]$ such that $\sum_{x \in X} \mu(x) = 1$. The set of all probability distributions over X is denoted by Δ_X.

2.1 Markov Decision Processes

We consider Markov decision processes (MDPs) with discounted rewards, defined as follows (Puterman 1994).

Definition 1 (MDP). *An MDP is a tuple $(S, s_I, A, P, R, \gamma)$, where S is a finite set of states, $s_I \in \Delta_S$ is the initial distribution, A is a finite set of actions, $P \colon S \times A \to \Delta_S$ is a transition function, $R \colon S \times A \to \mathbb{R}_{\geq 0}$ is a state-action reward function, and $\gamma \in (0, 1)$ is a discount factor.*

The actions in an MDP are chosen by a (randomized) policy $\pi \colon S \to \Delta_A$. We write Π for the set of all policies and simplify $\pi(s)(a)$ as $\pi(s, a)$. The objective in an MDP is to compute a policy π that maximizes the expected return ρ_P^π:

$$\rho_P^\pi \coloneqq \sum_{s \in S} s_I(s)V_P^\pi(s) = \langle s_I, V_P^\pi \rangle, \qquad (1)$$

where the value function $V_P^\pi \colon S \to \mathbb{R}$ is defined as

$$V_P^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^\pi(s_t) \,\Big|\, s_0 = s, s_{t+1} \sim P^\pi(s_t)\right],$$

with the transition and rewards functions for π given as

$$P^\pi(s) \coloneqq \sum_{a \in A} \pi(s, a)P(s, a) \in \Delta_S, \qquad (2)$$

$$R^\pi(s) \coloneqq \sum_{a \in A} \pi(s, a)R(s, a) \in \mathbb{R}_{\geq 0}. \qquad (3)$$

This value function is the fixed point of the Bellman operator $\mathcal{T}_P^\pi$ (Puterman 1994), which is defined for all states $s \in S$ as

$$(\mathcal{T}_P^\pi V)(s) \coloneqq [R^\pi(s) + \langle \gamma P^\pi(s), V \rangle],$$

whereas the optimal value $V_P^\star \coloneqq \max_{\pi \in \Pi} V_P^\pi$ is the fixed point of the optimal Bellman operator $\mathcal{T}_P^\star$ defined as

$$(\mathcal{T}_P^\star V)(s) \coloneqq \max_{\pi \in \Pi} \mathcal{T}_P^\pi V(s).$$

Thus, the sequences $V_{n+1}^\pi \coloneqq \mathcal{T}_P^\pi V_n^\pi$ and $V_{n+1}^\star \coloneqq \mathcal{T}_P^\star V_n^\star$ converge to their respective fixed points, *i.e.*, $\lim_{n\to\infty} V_n^\pi = V_P^\pi$ and $\lim_{n\to\infty} V_n^\star = V_P^\star$. Subsequently, an optimal policy can be computed as $\pi_P^\star \in \operatorname{argmax}_{\pi \in \Pi} \mathcal{T}_P^\pi V_P^\star$.

2.2 Robust Markov Decision Processes

Robust MDPs (RMDPs) extend MDPs with *sets of transition probabilities* (Iyengar 2005; Nilim and Ghaoui 2005). In an RMDP, the transition function is chosen from a set $\mathcal{P} \subseteq \{P \colon S \times A \to \Delta_S\}$ of transition functions, called the *uncertainty set* (also known as the *ambiguity set*).

Definition 2 (RMDP). *A robust MDP (RMDP) is a tuple $(S, s_I, A, \mathcal{P}, R, \gamma)$, where S, s_I, A, R, and γ are defined as in an MDP, and where $\mathcal{P} \subseteq \{P \colon S \times A \to \Delta_S\}$ is a set of transition functions, called the* uncertainty set.

The robust expected return $\rho_{\mathcal{P}}^\pi$ for the policy π is defined as the worst-case expected return over the uncertainty set $\mathcal{P}$:

$$\rho_{\mathcal{P}}^\pi \coloneqq \min_{P \in \mathcal{P}} \rho_P^\pi. \qquad (4)$$

The standard objective in an RMDP is to find an *optimal robust policy* $\pi_{\mathcal{P}}^\star$ maximizing the robust expected return $\rho_{\mathcal{P}}^\star$:

$$\pi_{\mathcal{P}}^\star \in \operatorname*{argmax}_{\pi \in \Pi} \rho_{\mathcal{P}}^\pi, \quad \rho_{\mathcal{P}}^\star \coloneqq \max_{\pi \in \Pi} \rho_{\mathcal{P}}^\pi. \qquad (5)$$

Unfortunately, solving Eqs. (4) and (5) is NP-hard for general uncertainty sets $\mathcal{P}$, even if they are convex (Wiesemann, Kuhn, and Rustem 2013). Thus, $\mathcal{P}$ is commonly assumed to be decomposable over states and/or state-action pairs, which is also known as *rectangularity* of the uncertainty set.

Definition 3 (Rectangularity). *The uncertainty set $\mathcal{P}$ is s-rectangular if it can be decomposed state-wise as $\mathcal{P} = \times_{s \in S} \mathcal{P}_s$, where $\mathcal{P}_s \subseteq \{P \colon Act \to \Delta_S\}$. Moreover, $\mathcal{P}$ is (s, a)-rectangular if it can be decomposed state-action-wise as $\mathcal{P} = \times_{s \in S, a \in A} \mathcal{P}_{s,a}$, where $\mathcal{P}_{s,a} \subseteq \Delta_S$.*

(s, a)-rectangularity is a special case of s-rectangularity.

Assumption 1. *Throughout the paper, the uncertainty set $\mathcal{P}$ of an RMDP is assumed to be s-rectangular.*

Under s-rectangularity, optimal policies may need to be randomized (Wiesemann, Kuhn, and Rustem 2013, Prop. 1). Our definitions follow the usual semantics that the environment knows the stochastic policy of the agent but not the actual actions sampled from this policy, known as the *environment first* (or *nature first*) semantics (Suilen et al. 2024).

Robust value iteration. Under s-rectangularity, for every policy π, there is a robust value function $V_{\mathcal{P}}^{\pi} : S \to \mathbb{R}$ that satisfies $V_{\mathcal{P}}^{\pi}(s) := \min_{P \in \mathcal{P}} V_P^{\pi}(s)$ for all $s \in S$ (Wiesemann, Kuhn, and Rustem 2013). This value function $V_{\mathcal{P}}^{\pi}$ is the fixed point of the robust Bellman operator $\mathcal{T}_{\mathcal{P}}^{\pi}$ for every $s \in S$:

$$(\mathcal{T}_{\mathcal{P}}^{\pi} V)(s) := \min_{P \in \mathcal{P}_S} \left[R^{\pi}(s) + \langle \gamma P^{\pi}(s), V \rangle \right].$$

Similarly, there exists an optimal robust value function $V_{\mathcal{P}}^{\star} := \max_{\pi \in \Pi} V_{\mathcal{P}}^{\pi}$, which is the fixed point of the optimal robust Bellman operator $\mathcal{T}_{\mathcal{P}}^{\star}$, defined for all $s \in S$ as

$$(\mathcal{T}_{\mathcal{P}}^{\star} V)(s) := \max_{\pi \in \Pi} \mathcal{T}_{\mathcal{P}}^{\pi} V(s).$$

Robust value iteration leverages these fixed points so that the sequences $V_{n+1}^{\pi} := \mathcal{T}_{\mathcal{P}}^{\pi} V_n^{\pi}$ and $V_{n+1}^{\star} := \mathcal{T}_{\mathcal{P}}^{\star} V_n^{\star}$ converge to their respective fixed points, *i.e.*, $\lim_{n \to \infty} V_n^{\pi} = V_{\mathcal{P}}^{\pi}$ and $\lim_{n \to \infty} V_n^{\star} = V_{\mathcal{P}}^{\star}$. Subsequently, an optimal robust policy can be computed as $\pi_{\mathcal{P}}^{\star} \in \mathrm{argmax}_{\pi} \mathcal{T}_{\mathcal{P}}^{\pi} V_{\mathcal{P}}^{\star}$.

3 Best-Effort Policies in RMDPs

The optimal robust policy in Eq. (5) assumes the choice of transition function from the uncertainty set to be fully adversarial. Here, we introduce *dominance* and *best-effort* as the basis for a policy selection criterion that also considers non-adversarial scenarios. These notions have been used in uncertain stochastic games (Aminof et al. 2023), but, as we discuss in Sect. 7, these results do not carry over to RMDPs.

3.1 Dominant and Best-Effort Policies

In this section, we tailor the definitions of *dominant* and *best-effort* policies from Aminof et al. (2023) to RMDPs. The first concept is that of *dominance* between policies.

Definition 4 (Dominance). *Let $\pi, \pi' \in \Pi$ be policies for the RMDP $\mathcal{M}_R$. The policy π dominates π', written $\pi \geq_{\mathcal{P}} \pi'$, if and only if $\rho_P^{\pi} \geq \rho_P^{\pi'}$ for all $P \in \mathcal{P}$.*

Intuitively, π dominates π' if π does not perform worse than π' under any transition function $P \in \mathcal{P}$. If, in addition, the policy π also attains a *strictly higher* expected return in some $P \in \mathcal{P}$, then π *strictly* dominates π':

Definition 5 (Strict dominance). *Let $\pi, \pi' \in \Pi$ be policies for RMDP $\mathcal{M}_R$. Policy π dominates π', written $\pi >_{\mathcal{P}} \pi'$, if and only if $\pi \geq_{\mathcal{P}} \pi'$ and there exists $P' \in \mathcal{P}$ s.t. $\rho_{P'}^{\pi} > \rho_{P'}^{\pi'}$.*

We say that the policy π is (strictly) dominant in the RMDP $\mathcal{M}_R$ if it (strictly) dominates every other policy $\pi' \in \Pi \setminus \{\pi\}$. Next, we say that a policy is *best-effort* if there is no other policy that dominates it.

Definition 6 (Best-effort). *A policy $\pi \in \Pi$ for the RMDP $\mathcal{M}_R$ is best-effort if there is no $\pi' \in \Pi$ such that $\pi' >_{\mathcal{P}} \pi$. We denote by $\Pi_{\mathrm{BE}} \subseteq \Pi$ the set of all best-effort policies.*

A policy is best-effort if there is no other policy that is strictly better for some $P \in \mathcal{P}$ and not worse for all $P \in \mathcal{P}$. In other words, a best-effort policy cannot be improved without also decreasing the expected return under some transition function. Best-effort policies are *incomparable* with respect to the dominance order, *i.e.*, for all $\pi, \pi' \in \Pi_{\mathrm{BE}}$, $\pi \neq \pi'$, we have both $\pi \not\geq_{\mathcal{P}} \pi'$ and $\pi \not\leq_{\mathcal{P}} \pi'$.
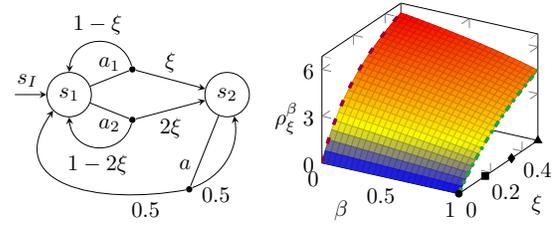

Figure 1: *Left:* An RMDP with two states, where the policy is fully defined by the probability $\beta := \pi(s_1, a_1)$ of choosing a_1 in s_1. The reward function is defined as $R(s_1, a_1) = R(s_1, a_2) = 0$ and $R(s_2, a) = 1$. *Right:* The expected return ρ_{ξ}^{β} as a function of β and $\xi \in [0, 0.5]$. All policies are optimal robust, but only the policy with $\beta = 0$ is best-effort.

3.2 Optimal Robust Does Not Imply Best-Effort

In general, optimal robust policies for RMDPs are not best-effort, as shown by the two-state RMDP in Fig. 1 with reward function $R(s_1, a_1) = R(s_1, a_2) = 0$, $R(s_2, a) = 1$. In this RMDP, the only action choice is between a_1 and a_2 in state s_1, so the stochastic policy π is completely defined by the probability $\beta := \pi(s_1, a_1) \in [0, 1]$ of choosing a_1 in s_1. Similarly, the (s-rectangular) uncertainty set $\mathcal{P}$ is fully defined by the parameter $\xi \in \Xi = [0, 0.5]$. As a result, we may simplify the notations from the preceding definitions by replacing π with β, P with ξ, and $\mathcal{P}$ with Ξ. The value function $V_{\xi}^{\beta} : \{s_1, s_2\} \to \mathbb{R}$ depends on β and ξ and is defined as

$$V_{\xi}^{\beta}(s_1) = \gamma \big(\beta \big[\xi V_{\xi}^{\beta}(s_2) + (1 - \xi) V_{\xi}^{\beta}(s_1) \big] \\ + (1 - \beta) \big[2\xi V_{\xi}^{\beta}(s_2) + (1 - 2\xi) V_{\xi}^{\beta}(s_1) \big] \big),$$

$$V_{\xi}^{\beta}(s_2) = 1 + \gamma \big[0.5 V_{\xi}^{\beta}(s_1) + 0.5 V_{\xi}^{\beta}(s_2) \big].$$

Solving for the value $\rho_{\xi}^{\beta} = V_{\xi}^{\beta}(s_1)$ with $\gamma = 0.9$ yields the surface in the right half of Fig. 1. This surface shows the expected return ρ_{ξ}^{β} for all $\beta \in [0, 1]$ and $\xi \in [0, 0.5]$. The worst-case expected return is zero and is attained at $\xi = 0$, regardless of the value of β. Thus, all policies in this RMDP are optimal robust. Nevertheless, only the policy for $\beta = 0$ is best-effort, because for all $\xi > 0$, the expected return for $\beta = 0$ is *strictly higher* than for any $\beta > 0$. In other words, the policy defined as $\pi(s_1, a_1) = 0$, $\pi(s_1, a_2) = 1$ *strictly dominates* all other policies $\pi' \neq \pi$, that is, $\pi >_{\Xi} \pi'$.

3.3 Problem Statement

Above, we have shown that not all optimal robust policies are also best-effort. This observation motivates the next core problem, which we shall solve in the remainder of this paper.

Problem 1. *For a given RMDP $\mathcal{M}_R$, compute a policy $\pi_{\mathrm{BE}}^{\star}$ that is optimal robust and best-effort, i.e.,*

$$\pi_{\mathrm{BE}}^{\star} \in \underset{\pi \in \Pi}{\mathrm{argmax}} \, \rho_{\mathcal{P}}^{\pi} \text{ such that } \nexists \pi' \in \Pi \setminus \{\pi_{\mathrm{BE}}^{\star}\}, \, \pi' >_{\mathcal{P}} \pi_{\mathrm{BE}}^{\star}.$$

We call a policy that satisfies Problem 1 *optimal robust best-effort* (ORBE). In RMDPs with multiple optimal robust policies (as in Fig. 1), the best-effort criterion offers a principled tiebreaker, favoring a policy that attains a maximal performance under non-adversarial transition probabilities.

Remark 1. *For clarity and due to space constraints, all proofs are provided in Appendix A.*

4 Representation of Robust Value Functions

We first introduce a change in perspective to the value function, which we will use in Sect. 5 to solve Problem 1. Instead of using shared variables to represent dependencies between probabilities (such as ξ in Fig. 1), we label each transition with its own probability $p(s, a)(s')$ and encode dependencies in the uncertainty set $\mathcal{P}$. For instance, we can equally represent the RMDP in Fig. 1 using the uncertainty set

$$\mathcal{P}_{s_1} = \{(P_{s_1} \colon A \to \Delta_S) : p(a_1)(s_1) + p(a_1)(s_2) = 1,$$
$$p(a_2)(s_1) + p(a_2)(s_2) = 1, \ p(a_1)(s_2) = 0.5 p(a_2)(s_2)\}.$$

We aim to reason about the expected return when the transition function is fixed *in all but one state*. To this end, we introduce the notion of a partial transition function.

Definition 7 (Partial transition function). *Let $\mathcal{P}$ be an s-rectangular uncertainty set and let $\bar{s} \in S$ be a state. A partial transition function $P_{-\bar{s}}$ for state $\bar{s}$ is defined as $P_{-\bar{s}} = \times_{s \in S \setminus \{\bar{s}\}} P_s$, where $P_s \in \mathcal{P}$ for all $s \in S \setminus \{\bar{s}\}$.*

A partial transition function has the form $P_{-\bar{s}} \colon (S \setminus \{\bar{s}\}) \times Act \to \Delta_S$. Thus, to complete $P_{-\bar{s}}$ with $P_{\bar{s}} \in \mathcal{P}_{\bar{s}}$ for the missing state $\bar{s}$, we take the product $P_{-\bar{s}} \times P_{\bar{s}}$. Similarly, we write $P_{-\bar{s}} \times \mathcal{P}_{\bar{s}}$ for the set of all completions, such that $P_{-\bar{s}} \times P_{\bar{s}} \in P_{-\bar{s}} \times \mathcal{P}_{\bar{s}}$. Using this notation, we define the following value function in a fixed state $\bar{s}$, when the transition probabilities are fixed in all states but $\bar{s}$.

Definition 8 (Parametric value function). *The value in state $\bar{s}$ is a function of the completion $P_{\bar{s}} \in \mathcal{P}_{\bar{s}}$ of the partial transition function $P_{-\bar{s}}$ and is defined as $Z^\pi_{P,\bar{s}}(P_{\bar{s}}) = V^\pi_{P_{-\bar{s}} \times P_{\bar{s}}}(\bar{s})$.*

Example 1. *Consider again the RMDP from Fig. 1 with the policies given by $\beta = 0$ and $\beta = 1$. The value functions Z^π_{P,s_1} for these two policies are, respectively, shown in the left and right halves of Fig. 2. For $\beta = 1$, the value depends only of the transition probabilities related to action a_1 (and for $\beta = 0$ only of those related to a_2). In both plots, the dashed line in the bottom plane shows the set of valid distributions in $\mathcal{P}_{s_1}$, where the marked points coincide with those on the ξ-axis in Fig. 1. The green (left) and purple (right) curved lines show the expected return for the policies with $\beta = 1$ and $\beta = 0$, respectively, as a function of ξ and coincide with the lines of the same color in Fig. 1. As in Fig. 1, we observe that, for any $\xi > 0$, the policy for $\beta = 1$ strictly dominates all other policies and is, thus, best-effort.*

5 Finding Optimal Robust Best-Effort Policies

We now use the representation of the value function from Def. 8 to determine whether an optimal robust policy for a fixed RMDP $\mathcal{M}_R = (S, s_I, A, \mathcal{P}, R, \gamma)$ is best-effort.

5.1 Existence of ORBE Policies

We first establish in Theorem 1 that, for any s-rectangular RMDP, the set of ORBE policies is nonempty. Intuitively, this result holds because the dominance relation imposes a partial order over policies, ensuring the existence of maximal
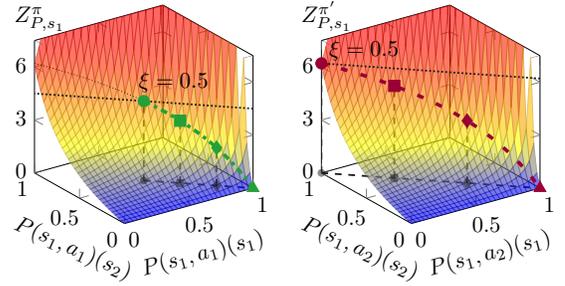

Figure 2: The value function Z^π_{P,s_1} in state s_1 for the RMDP from Fig. 1, shown for the policies with $\beta = 1$ (left) and $\beta = 0$ (right). The curved lines show the expected return as the parameter ξ in Fig. 1 ranges from 0 to 0.5 (the line markers correspond with those on the ξ-axis in Fig. 1).

(*i.e.*, best-effort) ones that under an adversarial environment must also be optimal robust. Furthermore, any optimal robust policy $\pi \in \Pi^\star$ cannot be dominated by a policy that is not optimal robust. Thus, an ORBE policy always exists.

Theorem 1 (Existence of ORBE policies). *For any RMDP, the intersection of the sets of* optimal robust policies $\Pi^\star$ *and* best-effort policies Π_{BE} *is nonempty.*

Note that the existence of best-effort policies in RMDPs does not directly follow from the results for synthesis in stochastic environments in Aminof et al. (2023); see Sect. 7.

5.2 Characterizing ORBE policies

In Theorem 2, we provide a sufficient condition for ORBE policies, used as a foundation in the remainder of the section.

Theorem 2 (ORBE policy). *Given an optimal robust policy $\pi^\star \in \Pi^\star := \operatorname{argmax}_{\pi \in \Pi} \rho^\pi_{\mathcal{P}}$, if there exists $P \in \mathcal{P}$ such that $\rho^{\pi^\star}_P > \rho^{\pi'}_P$ for all $\pi' \in \Pi^\star \setminus \{\pi^\star\}$, then $\pi^\star$ is ORBE.*

Proof. First, $\pi^\star$ is optimal robust by definition. Second, to show that $\pi^\star$ is also best-effort, we must show there is no other policy $\pi' \in \Pi \setminus \{\pi^\star\}$ that strictly dominates $\pi^\star$. By construction, $\pi^\star$ cannot be dominated by any $\pi' \in \Pi^\star \setminus \{\pi^\star\}$. For any other policy $\pi'' \in \Pi \setminus \Pi^\star$, we have $\rho^{\pi^\star}_{\mathcal{P}} > \rho^{\pi''}_{\mathcal{P}}$ and, moreover, as $\rho^{\pi^\star}_{\mathcal{P}} = \min_{P \in \mathcal{P}} \rho^\pi_P$ (cf. Eq. (4)), it holds that $\rho^{\pi^\star}_{P'} \geq \rho^{\pi^\star}_{\mathcal{P}}$ for all $P' \in \mathcal{P}$. By letting $P' \in \operatorname{argmin}_{P \in \mathcal{P}} \rho^{\pi''}_P$, we thus obtain $\rho^{\pi^\star}_{P'} \geq \rho^{\pi^\star}_{\mathcal{P}} > \rho^{\pi''}_{\mathcal{P}} = \rho^{\pi''}_{P'}$, which proves that $\pi'' \not\succ_{\mathcal{P}} \pi^\star$. Thus, the policy $\pi^\star$ is ORBE. $\square$

In the remainder of this section, we use Theorem 2 to derive conditions under which an optimal robust policy is also best-effort (and thus ORBE). First of all, if an optimal robust policy is unique, then this policy is also best-effort.

Corollary 1. *Let $\Pi^\star = \operatorname{argmax}_{\pi \in \Pi} \rho^\pi_{\mathcal{P}}$ be the set of optimal robust policies. If $\Pi^\star$ is a singleton, then $\pi^\star \in \Pi^\star$ is ORBE.*

ORBE via optimistic RVI. The second observation is that, if an optimal robust policy is not unique but further optimizing via robust value iteration (RVI) for the *optimistic* (*i.e.*, maximizing) transition function does yield a unique optimum, then the resulting policy is also best-effort.
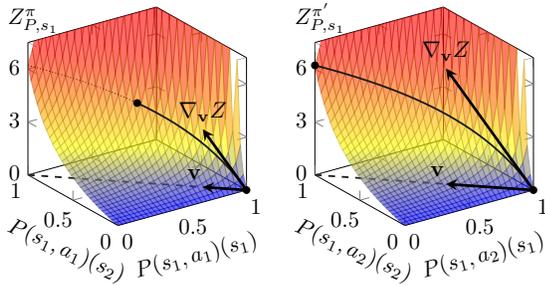
Figure 3: The directional derivative $\nabla_{\mathbf{v}} Z_{P,s_1}^{\pi}$ for $\beta = 0$ (shown in the right half) is strictly larger than for any $\beta > 0$. Hence, we conclude that the policy for $\beta = 0$ is ORBE.

Corollary 2. *Let* $\check{\Pi}^{\star} = \operatorname{argmax}_{\pi \in \Pi} \rho_{\mathcal{P}}^{\pi}$ *and let* $\hat{\Pi}^{\star} = \operatorname{argmax}_{\pi \in \check{\Pi}^{\star}} \max_{P \in \mathcal{P}} \rho_P^{\pi}$ *be the set of policies that (within* $\check{\Pi}^{\star}$*) maximize the expected return under the maximizing* $P \in \mathcal{P}$*. If* $\hat{\Pi}^{\star}$ *is a singleton, then* $\pi^{\star} \in \hat{\Pi}^{\star}$ *is* ORBE.

Example 2. *Consider again the RMDP in Fig. 1. Even though all policies are optimal robust, only the policy for* $\beta = 0$ *is optimal under the maximizing transition function (which is attained for* $\xi = 0.5$*). Thus, the policy for* $\beta = 0$*, i.e., always choosing action* a_2*, is* ORBE.

ORBE via derivatives. Another way to determine if a policy is best-effort is to reason about the derivative of the value function. Let $\nabla_{\mathbf{v}} f(x) = \mathbf{v}^{\top} \cdot \frac{\partial f(x)}{\partial x}$ be the *directional derivative* of the function $f : \mathbb{R}^n \to \mathbb{R}$ in the direction $\mathbf{v} \in \mathbb{R}^n$. Recall from Def. 8 that $Z_{P,\bar{s}}$ is the value function in state $\bar{s}$ when the transition function is fixed in all states except $\bar{s}$. The next result states that, if an optimal robust policy $\pi^{\star}$ leads, for every state $\bar{s}$, to a *strictly higher* derivative of $Z_{P,\bar{s}}^{\pi}(P_{\bar{s}})$ than all other optimal robust policies, then $\pi^{\star}$ is best-effort. This derivative can be taken in any direction such that the perturbed $P_{\bar{s}}$ is still within the uncertainty set $\mathcal{P}_{\bar{s}}$.

Corollary 3. *Let* $\bar{\pi} \in \Pi^{\star} = \operatorname{argmax}_{\pi \in \Pi} \rho_{\mathcal{P}}^{\bar{\pi}}$ *be an optimal robust policy with minimizer* $P^{\star} \in \operatorname{argmin}_{P \in \mathcal{P}} \rho_P^{\bar{\pi}}$*. Define* $\Pi_{(1)}^{\star} = \operatorname{argmax}_{\pi \in \Pi^{\star}} \rho_{P^{\star}}^{\pi}$ *and pick a policy* $\pi^{\star} \in \Pi_{(1)}^{\star}$*. The policy* $\pi^{\star}$ *is* ORBE *if, for all states* $\bar{s}$*, there exists a vector* $\mathbf{v} \in \mathbb{R}^{|S|}$ *such that* $\exists \epsilon > 0, P_{\bar{s}}^{\star} + \epsilon \mathbf{v} \in \mathcal{P}_{\bar{s}}$ *and*

$$\nabla_{\mathbf{v}} Z_{P^{\star},\bar{s}}^{\pi^{\star}}(P_{\bar{s}}^{\star}) > \nabla_{\mathbf{v}} Z_{P^{\star},\bar{s}}^{\pi'}(P_{\bar{s}}^{\star}) \ \forall \pi' \in \Pi_{(1)}^{\star} \setminus \{\pi^{\star}\}. \quad (6)$$

Intuitively, the condition that there exists $\epsilon > 0$ such that $P_{\bar{s}}^{\star} + \epsilon \mathbf{v} \in \mathcal{P}_{\bar{s}}$ encodes that the vector $\mathbf{v}$ at the minimizing transition function $P^{\star}$ points inside the uncertainty set $\mathcal{P}_{\bar{s}}$.

Example 3. *Another way to characterize the* ORBE *policy of* $\beta = 0$ *in Fig. 1 is to compare the derivatives of* Z_{P,s_1}^{π} *at the minimizing transition function* $\xi = 0$*. In this example, the only feasible direction is* $\mathbf{v} = [-\alpha, \alpha]$*,* $\alpha > 0$*, as shown in Fig. 3. Any change in* $P(s_1, a_1)(s_1)$ *and* $P(s_1, a_1)(s_2)$ *causes a change twice as big in* $P(s_1, a_2)(s_1)$ *and* $P(s_1, a_2)(s_1)$*, visualized by the longer vectors in Fig. 3. Thus, the directional derivative for the policy with* $\beta = 0$ *(i.e., always choosing action* a_2*) is strictly larger than for all* $\beta > 0$*. Therefore, the policy with* $\beta = 0$ *is* ORBE.

Conversely, we can consider the derivative of the value function under the policies $\pi^{\star} \in \hat{\Pi}^{\star}$ (as defined in Corollary 2) that are, besides being optimal robust, also optimal under the maximizing transition function. In this case, if $\pi^{\star}$ leads to a strictly *lower* directional derivative $\nabla_{\mathbf{v}} Z_{P,\bar{s}}^{\pi^{\star}}(P_{\bar{s}})$ than all other policies $\pi' \in \Pi^{\star} \setminus \{\pi^{\star}\}$, then $\pi^{\star}$ is best-effort. This result and the proof are analogous to Corollary 3, so we omit a formal statement due to space limitations.

Completeness. So far, we have shown that an optimal robust policy $\pi^{\star}$ is best-effort if either of the following holds:
1. $\pi^{\star}$ is *uniquely* optimal (in the minimizing or maximizing sense with respect to the transition function).
2. $\pi^{\star}$ yields a *uniquely* highest (resp. lowest) directional derivative at the minimizing (resp. maximizing) $P^{\star} \in \mathcal{P}$.

In this section, we complete the characterization by showing that any policy that satisfies these conditions up to this uniqueness is also best-effort. Theorem 3 formalizes this non-trivial result. For conciseness, we defer the preliminaries needed for the proof of Theorem 3 to Appendix A.4.

Theorem 3 (Computing ORBE policies). *Let* $\Pi^{\star} = \operatorname{argmax}_{\pi \in \Pi} \rho_{\mathcal{P}}^{\pi}$ *be the optimal robust policies. Pick two transition functions* $P^{(1)}, P^{(2)} \in \mathcal{P}$ *such that* $P^{(1)} \neq P^{(2)}$ *and, for all* $\bar{s} \in S$*, the line* $g_{\bar{s}}(\lambda) = \lambda P_{\bar{s}}^{(1)} + (1-\lambda) P_{\bar{s}}^{(2)}$ *intersects the relative interior[1] of* $\mathcal{P}_{\bar{s}}$*, or* $\mathcal{P}_{\bar{s}} \cap \{g_{\bar{s}}(\lambda)\} = \mathcal{P}_{\bar{s}}$*. Define*

$$\Pi_{(1)}^{\star} = \operatorname*{argmax}_{\pi \in \Pi^{\star}} \rho_{P^{(1)}}^{\pi}, \qquad \Pi_{(2)}^{\star} = \operatorname*{argmax}_{\pi \in \Pi_{(1)}^{\star}} \rho_{P^{(2)}}^{\pi}.$$

Choose a policy $\pi^{\star} \in \Pi_{(2)}^{\star}$ *s.t., for all* $\bar{s} \in S$*, it holds that*

$$\nabla_{\mathbf{v}} Z_{P^{(1)},\bar{s}}^{\pi^{\star}}(P_{\bar{s}}^{(1)}) \geq \nabla_{\mathbf{v}} Z_{P^{(1)},\bar{s}}^{\pi'}(P_{\bar{s}}^{(1)}) \ \forall \pi' \in \Pi_{(2)}^{\star}, \quad (7\text{a})$$

$$\nabla_{\mathbf{v}} Z_{P^{(2)},\bar{s}}^{\pi^{\star}}(P_{\bar{s}}^{(2)}) \leq \nabla_{\mathbf{v}} Z_{P^{(2)},\bar{s}}^{\pi'}(P_{\bar{s}}^{(2)}) \ \forall \pi' \in \Pi_{(2)}^{\star}, \quad (7\text{b})$$

where the vector $\mathbf{v} \in \mathbb{R}^{|S|}$ *is defined as*

$$\mathbf{v} = \begin{cases} P^{(2)} - P^{(1)} & \text{if } \rho_{P^{(2)}}^{\pi^{\star}} > \rho_{P^{(1)}}^{\pi^{\star}}, \\ P^{(1)} - P^{(2)} & \text{otherwise.} \end{cases}$$

Then, the policy $\pi^{\star}$ *is* ORBE

In the proof, presented in Appendix A.4, we show that there always exists a policy $\pi^{\star} \in \Pi_{(2)}^{\star}$ that satisfies Eqs. (7a) and (7b). As discussed next, a practical implementation of Theorem 3 is to choose $P^{(1)}$ and $P^{(2)}$ as worst- and best-case transition functions.

5.3 Algorithm
Theorem 3 leads to Algorithm 1 for computing an ORBE policy. In particular we iteratively refine Π by applying the criteria presented above to obtain an ORBE policy.

We first use robust value iteration to compute the set of optimal robust policies (Line 1), which, if a singleton[2], consists

---

[1] The relative interior of a convex set X is defined as $\operatorname{relint}(X) := \{x \in X : \forall y \in X, \exists \lambda > 1. \ \lambda x + (1-\lambda)y \in X\}$.

[2] An optimal policy $\pi^{\star}$ is unique if, for every state $s \in S$, the robust value $V_{\mathcal{P}}^{\pi^{\star}}(s)$ is strictly higher than $R(s,a) + \langle \gamma P(s,a), V_{\mathcal{P}}^{\pi^{\star}} \rangle$ for all other actions $a \neq \pi^{\star}(s)$ (Puterman 1994). For randomized policies, we instead must check for strict concavity of the value function with respect to the policy, *e.g.*, by deriving the optimal robust Bellman operator explicitly as in Kumar et al. (2024).

Algorithm 1: Computation of ORBE policy.

Input: s-rectangular RMDP $(S, A, \mathcal{P}, r, \gamma)$
Output: ORBE policy $\pi^\star \in \Pi^\star_{\mathrm{BE}}$

1: $\Pi \leftarrow \mathrm{argmax}_\pi \min_{P \in \mathcal{P}} \rho_P^\pi$
2: **if** $|\Pi| > 1$ **then**
3:     $\Pi \leftarrow \mathrm{argmax}_{\pi \in \Pi} \max_{P \in \mathcal{P}} \rho_P^\pi$
4:     **if** $|\Pi| > 1$ **then**
5:        $\pi \leftarrow \Pi$
6:        $P^{(1)} \leftarrow \mathrm{argmin}_{P \in \mathcal{P}} \rho_P^\pi$
7:        $P^{(2)} \leftarrow \mathrm{argmax}_{P \in \mathcal{P}} \rho_P^\pi$
8:        $\Pi \leftarrow \mathrm{argmax}_{\pi \in \Pi} \rho_{P^{(1)}}^\pi$
9:        $\Pi \leftarrow \mathrm{argmax}_{\pi \in \Pi} \rho_{P^{(2)}}^\pi$
10:       $\mathbf{v} \leftarrow P^{(2)} - P^{(1)} \;\; \forall \bar{s}$
11:       $\Pi \leftarrow \times_{\bar{s} \in S} \, \mathrm{argmax}_{\pi(\bar{s}) \in \Pi} \, \nabla_{\mathbf{v}} Z^\pi_{P^{(1)}, \bar{s}}(P^{(1)}_{\bar{s}})$
12:       **if** $|\Pi| > 1$ **then**
13:          $\Pi \leftarrow \times_{\bar{s} \in S} \, \mathrm{argmin}_{\pi(\bar{s}) \in \Pi} \, \nabla_{\mathbf{v}} Z^\pi_{P^{(2)}, \bar{s}}(P^{(2)}_{\bar{s}})$
14: **return** any $\pi^\star \in \Pi$

---

of an ORBE policy by Corollary 1, thus solving Problem 1. Otherwise, we analogously compute the set of optimal policies under the maximizing transition function (Line 3), which, if a singleton, contains an ORBE policy by Corollary 2.

If this set is still not a singleton, we arbitrarily select a policy π from the remaining updated set Π (Line 5) and compute the minimizing and maximizing transition functions $P^{(1)}$ and $P^{(2)}$ (Lines 6 and 7). We then refine the policy set by keeping only those that first maximize the expected return for $P^{(1)}$ and then for $P^{(2)}$ (Lines 8 and 9). For every $\bar{s} \in S$, we define $\mathbf{v} \leftarrow P^{(2)} - P^{(1)}$ as per Theorem 3 (Line 10). Next, we refine the set of policies by, in every state $\bar{s} \in S$, only selecting actions that *maximize* the directional derivative at the *minimizer* $P^{(1)}$ (Line 11). The Cartesian product $\Pi \leftarrow \times_{\bar{s} \in S} \cdots$ of these actions gives the set of policies that satisfy Eq. (7a). If Π is now a singleton, then it satisfies Corollary 3 and, thus, $\pi^\star \in \Pi$ is ORBE. Otherwise, if multiple policies remain, we perform the analogous refinement—over the set of policies obtained in Line 11—to *minimize* the directional derivative at the *maximizing* transition function $P^{(2)}$ (Line 13).

Any returned policy $\pi^\star$ satisfies at least one of the Corollaries 1 to 3 or Theorem 3, thus showing that the algorithm always returns a ORBE policy.

Remark 2. *We can easily amend Algorithm 1 for a policy that* minimizes *expected return under the* maximizing *probabilities. In this case, we replace all* min *with* max *and vice versa. We shall see such an application in Sect. 6.*

Complexity. The computations in Algorithm 1 lead to a manageable overhead compared to the standard robust value iteration in Line 2. First, Line 3 amounts to running robust value iteration again, but over a potentially smaller subset of actions per state, increasing complexity by a constant smaller than 2. Next, Lines 6 to 9 compute the minimizer and maximizer, and solve the two associated MDPs using standard value iteration. Finally, maximizing the derivatives (Line 11) amounts to solving a linear equation system of size $|S|$ for ev-

ery state and action (Heck et al. 2022; Badings et al. 2023a). Solving each equation system has worst-case complexity $O(|S|^3)$, yielding an overall complexity of $O(|S|^4 \cdot |A|)$ for Line 11 (and, by symmetry, also for Line 13). Thus, whenever computing an optimal robust policy is feasible, the additional overhead of Algorithm 1 is also manageable.

6 Empirical Evaluation

In Sect. 5, we presented an efficient and complete algorithm for computing ORBE policies. In this section, we experimentally show the applicability of our algorithm within different implementations of robust value iteration. Our primary objective is to provide a proof of concept to confirm the theoretical results from Sect. 5. The experiments ran on an Apple MacBook with an M4 Pro chip and 24GB of RAM. The code is available on https://github.com/tbadings/best-effort-rmdps.

6.1 Best-Effort Policies for Interval MDPs

We consider robust value iteration within PRISM, a popular tool for MDPs (Kwiatkowska, Norman, and Parker 2011). PRISM only supports *interval MDPs* (IMDPs), *i.e.*, (s, a)-rectangular RMDPs with interval-valued probabilities. We consider variants of a *slippery gridworld* IMDP (see Appendix B for details). The objective is to minimize the expected number of steps to reach the goal state. When the agent slips, it remains in the same state. The agent can move in each direction with two actions: one where the slipping probability p is *fixed*, and one where it belongs to the *interval* $[q, p]$. Since the goal is to minimize the number of steps, the worst-case slipping probability is p, so the robust value of both action types is the same. However, only a policy that always picks the interval-valued action is best-effort.

To show that PRISM returns an arbitrary optimal robust (but not necessarily ORBE) policy, we define the IMDP's actions in different orders. Let $\nu \in [0, 1]$ be the fraction of states in which the best-effort action is defined first. We consider $\nu = 0$ (non-best-effort always defined first), $\nu = 1$ (best-effort defined first), and $\nu = 0.5$ (a coin-flip decides which action is defined first). We repeat each experiment over 10 seeds. The results in Table 1 show the percentage of states where the optimal robust policy returned by PRISM chooses the best-effort action (*i.e.*, the action with an interval for the slipping probability). Essentially, the PRISM policy sticks to the first action it finds to be optimal robust, so the fraction of best-effort actions is roughly proportional to ν. Thus, PRISM finds optimal robust policies, but not necessarily ORBE ones.

Conversely, for our method, we apply Corollary 2 by again running robust value iteration with PRISM, but this time over the optimal robust policies and for the *best-case* slipping probability. This second run of value iteration is over a smaller set of policies and less than doubles the runtime (especially for $|S| = 10^4$), thus confirming our results from Sect. 5: the complexity for computing ORBE policies is still dominated by that of robust value iteration, making the process feasible whenever robust optimal policies can be computed. The policy obtained using our approach always chooses actions with the interval-valued slipping probabilities. Thus, and as confirmed by the rightmost column of Table 1, the use of Corollary 2 indeed always leads to ORBE policies.

$|S|$	ν	PRISM		+ Best-case (Corr. 2)	
		Time [s]	BE [%]	Time [s]	BE [%]
100	0.0	2.0	21.9	3.9	100.0
	0.5	1.9	59.6	3.8	100.0
	1.0	1.9	89.9	3.9	100.0
900	0.0	2.1	23.3	4.0	100.0
	0.5	2.1	62.0	4.1	100.0
	1.0	2.1	87.4	4.2	100.0
10 000	0.0	48.9	21.2	54.4	100.0
	0.5	54.9	39.3	61.2	100.0
	1.0	51.0	85.4	56.5	100.0

Table 1: Comparison to PRISM on the gridworld IMDPs, showing the grid sizes, probability ν to define the best-effort action first, runtimes, and percentage of states in which the resulting optimal policy chooses a best-effort (BE) action.

$|S|$	ν	RVI		+ Best-case (Corr. 2)		+ Deriv. (Corr. 3)	
		Time [s]	BE [%]	Time [s]	BE [%]	Time [s]	BE [%]
100	0.0	7.0	0.0	11.7	100.0	7.1	100.0
	0.5	7.0	49.2	11.7	100.0	7.1	100.0
	1.0	7.6	100.0	12.7	100.0	7.6	100.0
400	0.0	49.5	0.0	83.4	100.0	50.1	100.0
	0.5	50.1	48.0	84.4	100.0	50.6	100.0
	1.0	48.4	100.0	81.7	100.0	48.9	100.0
900	0.0	163.6	0.0	274.4	100.0	172.0	100.0
	0.5	163.4	50.1	273.8	100.0	171.9	100.0
	1.0	164.1	100.0	275.0	100.0	172.6	100.0

Table 2: Results on the gridworld RMDPs, for robust value iteration (RVI), RVI plus optimizing for the best-case probabilities, and RVI plus optimizing for the derivatives.

6.2 Best-Effort Policies for s-Rectangular RMDPs

To show the applicability of our methods beyond IMDPs, we create a basic implementation of robust value iteration and the derivative computation for s-rectangular RMDPs (see Appendix B for details). We consider variants of the same slippery gridworld as in Sect. 6 but now with an s-rectangular uncertainty set. For this RMDP, either Corollary 2 or 3 is sufficient to obtain an ORBE policy. Therefore, instead of implementing Algorithm 1 sequentially, we test both separately on top of robust value iteration.

The results in Table 2 give the same picture as in Sect. 6: if multiple optimal robust policies exist, robust value iteration returns the first optimal actions it finds. By contrast, our methods provide simple yet effective tie-break rules, either by returning a policy that is also optimal under the best-case transition probabilities (RVI + Corollary 2), or by returning a policy with the highest derivatives (RVI + Corollary 3). The former less than doubles the total runtime (especially for the larger models), while computing derivatives is even cheaper, increasing the total runtime by less than 10%.

7 Related Work

The notion of best-effort was first introduced in a game theoretic context by Faella (2009) as a relaxation of "winning" policies (or strategies). These ideas have been adapted to *reactive synthesis*, where in the absence of a winning strategy, best-effort policies can be computed at the same cost (Aminof et al. 2019, 2020; De Giacomo, Parretti, and Zhu 2025). Closest to our work are Aminof et al. (2023) and Giacomo, Favorito, and Silo (2024), who study best-effort for stochastic games where each transition probability is only constrained to lie within the open interval $(0, 1)$. Crucially, Aminof et al. (2023); Giacomo, Favorito, and Silo (2024) exploit this lack of probability bounds to construct a three-valued abstraction of policies (*winning*, *losing*, and *pending*) which is central to their characterization of best-effort policies. However, this does not carry over to RMDPs, where probabilities are bounded subsets of $[0, 1]$, thus breaking a direct translation of their characterization to the RMDP setting.

Related are *lexicographic orderings* over objectives for MDPs (Wray, Zilberstein, and Mouaddib 2015) and algorithms for stochastic games that progressively prune suboptimal actions per objective (Chatterjee et al. 2024). While our algorithm is conceptually similar, the refinement to best-effort policies requires different reasoning over the dominance order over policies. In *multi-objective MDPs* (MOMDPs), multiple objectives are combined, leading to Pareto optimality (Delgrange et al. 2020; Etessami et al. 2008). While MOMDPs require a trade-off between the objectives, our setting uses best-effort as a hard refinement within the optimal robust policies. Finally, weakly related are partial orders over states of MDPs (Roux and Pérez 2018) and monotonicity in parametric Markov chains (Spel, Junges, and Katoen 2019).

While we focus on s-rectangular RMDPs, our definitions of best-effort and dominance carry over to other models, such as k- or non-rectangular RMDPs (Mannor et al. 2004; Goyal and Grand-Clément 2023; Gadot et al. 2024) and parametric MDPs (Quatmann et al. 2016). However, computing optimal policies for these models is much harder—up to NP-hard for general non-rectangular RMDPs (Wiesemann, Kuhn, and Rustem 2013). Thus, adapting dynamic programming methods to these models is still an open problem.

8 Conclusion

We presented a principled tie-breaker among optimal robust policies in RMDPs based on best-effort. Our proposed ORBE policies maximize the worst-case expected return but also achieve a maximal expected return under non-adversarial transition probabilities. We fully characterized ORBE policies and presented an algorithm for computing them. Our experiments showed how to use our methods as an effective and efficient tie-breaker within robust value iteration.

Future work includes generalizing our methods to non-rectangular RMDPs or parametric MDPs. Moreover, our methods still rely on first computing a policy under adversarial transition probabilities. A next step is to consider ε-close optimal robust policies and optimize for best-effort within this broader context. Finally, we aim to study settings with a Bayesian prior over the uncertainty set (Murphy 2001).

Acknowledgments

References

Aminof, B.; Giacomo, G. D.; Lomuscio, A.; Murano, A.; and Rubin, S. 2020. Synthesizing strategies under expected and exceptional environment behaviors. In *IJCAI*, 1674–1680. ijcai.org.

Aminof, B.; Giacomo, G. D.; Murano, A.; and Rubin, S. 2019. Planning under LTL Environment Specifications. In *ICAPS*, 31–39. AAAI Press.

Aminof, B.; Giacomo, G. D.; Rubin, S.; and Zuleger, F. 2023. Stochastic Best-Effort Strategies for Borel Goals. In *LICS*, 1–13. IEEE.

Åström, K. J. 2012. *Introduction to stochastic control theory*. Courier Corporation.

Badings, T. S.; Junges, S.; Marandi, A.; Topcu, U.; and Jansen, N. 2023a. Efficient Sensitivity Analysis for Parametric Robust Markov Chains. In *CAV (3)*, volume 13966 of *Lecture Notes in Computer Science*, 62–85. Springer.

Badings, T. S.; Simão, T. D.; Suilen, M.; and Jansen, N. 2023b. Decision-making under uncertainty: beyond probabilities. *Int. J. Softw. Tools Technol. Transf.*, 25(3): 375–391.

Chatterjee, K.; Katoen, J.; Mohr, S.; Weininger, M.; and Winkler, T. 2024. Stochastic games with lexicographic objectives. *Formal Methods Syst. Des.*, 63(1): 40–80.

Davis, M. H. 2018. *Markov models & optimization*. Routledge.

De Giacomo, G.; Parretti, G.; and Zhu, S. 2025. Symbolic LTL$_f$ Synthesis: A Unified Approach for Synthesizing Winning, Dominant, and Best-Effort Strategies. *SN Computer Science*, 6(2): 147.

Delgrange, F.; Katoen, J.; Quatmann, T.; and Randour, M. 2020. Simple Strategies in Multi-Objective MDPs. In *TACAS (1)*, volume 12078 of *Lecture Notes in Computer Science*, 346–364. Springer.

Etessami, K.; Kwiatkowska, M. Z.; Vardi, M. Y.; and Yannakakis, M. 2008. Multi-Objective Model Checking of Markov Decision Processes. *Log. Methods Comput. Sci.*, 4(4).

Faella, M. 2009. Admissible Strategies in Infinite Games over Graphs. In *MFCS*, volume 5734 of *Lecture Notes in Computer Science*, 307–318. Springer.

Gadot, U.; Derman, E.; Kumar, N.; Elfatihi, M. M.; Levy, K.; and Mannor, S. 2024. Solving Non-rectangular Reward-Robust MDPs via Frequency Regularization. In *AAAI*, 21090–21098. AAAI Press.

Giacomo, G. D.; Favorito, M.; and Silo, L. 2024. Composition of Stochastic Services for LTL$_f$ Goal Specifications. In *FoIKS*, volume 14589 of *Lecture Notes in Computer Science*, 298–316. Springer.

Golub, G. H.; and Loan, C. F. V. 2013. *Matrix Computations, Fourth Edition*. Johns Hopkins University Press.

Goyal, V.; and Grand-Clément, J. 2023. Robust Markov Decision Processes: Beyond Rectangularity. *Math. Oper. Res.*, 48(1): 203–226.

Halmos, P. R. 1974. *Zorn's Lemma*, 62–65. New York, NY: Springer New York. ISBN 978-1-4757-1645-0.

Hanheide, M.; Göbelbecker, M.; Horn, G. S.; Pronobis, A.; Sjöö, K.; Aydemir, A.; Jensfelt, P.; Gretton, C.; Dearden, R.; Janícek, M.; Zender, H.; Kruijff, G. M.; Hawes, N.; and Wyatt, J. L. 2017. Robot task planning and explanation in open and uncertain worlds. *Artif. Intell.*, 247: 119–150.

Heck, L.; Spel, J.; Junges, S.; Moerman, J.; and Katoen, J. 2022. Gradient-Descent for Randomized Controllers Under Partial Observability. In *VMCAI*, volume 13182 of *Lecture Notes in Computer Science*, 127–150. Springer.

Iyengar, G. N. 2005. Robust Dynamic Programming. *Math. Oper. Res.*, 30(2): 257–280.

Junges, S.; Ábrahám, E.; Hensel, C.; Jansen, N.; Katoen, J.; Quatmann, T.; and Volk, M. 2024. Parameter synthesis for Markov models: covering the parameter space. *Formal Methods Syst. Des.*, 62(1): 181–259.

Kumar, N.; Wang, K.; Levy, K. Y.; and Mannor, S. 2024. Efficient Value Iteration for s-rectangular Robust Markov Decision Processes. In *ICML*. OpenReview.net.

Kwiatkowska, M. Z.; Norman, G.; and Parker, D. 2011. PRISM 4.0: Verification of Probabilistic Real-Time Systems. In *CAV*, volume 6806 of *Lecture Notes in Computer Science*, 585–591. Springer.

Mannor, S.; Simester, D.; Sun, P.; and Tsitsiklis, J. N. 2004. Bias and variance in value function estimation. In *ICML*, volume 69 of *ACM International Conference Proceeding Series*. ACM.

Moerland, T. M.; Broekens, J.; Plaat, A.; and Jonker, C. M. 2023. Model-based Reinforcement Learning: A Survey. *Found. Trends Mach. Learn.*, 16(1): 1–118.

Murphy, K. 2001. An introduction to graphical models. *Rap. tech*, 96: 1–19.

Nilim, A.; and Ghaoui, L. E. 2005. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Oper. Res.*, 53(5): 780–798.

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

Quatmann, T.; Dehnert, C.; Jansen, N.; Junges, S.; and Katoen, J. 2016. Parameter Synthesis for Markov Models: Faster Than Ever. In *ATVA*, volume 9938 of *Lecture Notes in Computer Science*, 50–67.

Roux, S. L.; and Pérez, G. A. 2018. The Complexity of Graph-Based Reductions for Reachability in Markov Decision Processes. In *FoSSaCS*, volume 10803 of *Lecture Notes in Computer Science*, 367–383. Springer.

Russell, S. J.; and Norvig, P. 2010. *Artificial Intelligence - A Modern Approach, Third International Edition*. Pearson Education.

Spel, J.; Junges, S.; and Katoen, J. 2019. Are Parametric Markov Chains Monotonic? In *ATVA*, volume 11781 of *Lecture Notes in Computer Science*, 479–496. Springer.

Suilen, M.; Badings, T. S.; Bovy, E. M.; Parker, D.; and Jansen, N. 2024. Robust Markov Decision Processes: A Place Where AI and Formal Methods Meet. In *Principles of Verification (3)*, volume 15262 of *Lecture Notes in Computer Science*, 126–154. Springer.

Wiesemann, W.; Kuhn, D.; and Rustem, B. 2013. Robust Markov Decision Processes. *Math. Oper. Res.*, 38(1): 153–183.

Wray, K. H.; Zilberstein, S.; and Mouaddib, A. 2015. Multi-Objective MDPs with Conditional Lexicographic Reward Preferences. In *AAAI*, 3418–3424. AAAI Press.

A Proofs

This appendix provides the complete proofs of the theoretical results presented in the main paper. For clarity and self-containment, all theorems and lemmas are restated before their corresponding proofs.

A.1 Existence of Best-Effort Policies

This section contains the proof of Theorem 1. We first prove that the set of best-effort policies in an s-rectangular RMDP is non-empty (Theorem 4) and then show that its intersection with the set of optimal robust is non-empty as well.

Theorem 4 (Existence of best-effort policies). *The set of best-effort policies* Π_{BE} *is nonempty.*

Proof. To prove the existence of at least one best-effort policy in an RMDP $\mathcal{M}$, we show the existence of a maximal element (with respect to the dominance order) in the set of policies Π in $\mathcal{M}$, *i.e.*, a best-effort policy.

We start by showing that the value functions induced by policies are smooth rational functions over the choice of transition probabilities (Puterman 1994, Theorem 6.1.1), and are, therefore, well-defined and comparable over the uncertainty set. The expected return $V_P^\pi \in \mathbb{R}^{|S|}$ under the policy $\pi \in \Pi$ and transition function $P \in \mathcal{P}$ is written in matrix form as

$$V_P^\pi = (I - \gamma P^\pi)^{-1} R^\pi, \tag{8}$$

where $P^\pi \in \mathbb{R}^{|S| \times |S|}$ and $R^\pi \in \mathbb{R}^{|S|}$ are the matrix and vector forms of Eqs. (2) and (3) over all states, respectively. Thus, the expected return $\rho_P^\pi = \langle s_I, V_P^\pi \rangle$ for a fixed policy π is a smooth rational function over the transition function $P \in \mathcal{P}$, which is called the *solution function*.[3] Note that, since the solution functions induced by policies are smooth, any strict improvement occurs over a nontrivial (non-measure-zero) subset of the domain.

Recall that the dominance relation $\geq_{\mathcal{P}}$ in Def. 5 is defined as follows:

Let $\pi, \pi' \in \Pi$ *be policies for the RMDP* $\mathcal{M}_R$. *The policy* π *dominates* π', *written* $\pi \geq_{\mathcal{P}} \pi'$, *if and only if* $\rho_P^\pi \geq \rho_P^{\pi'}$ *for all* $P \in \mathcal{P}$.

This relation is reflexive, transitive, and antisymmetric. Thus, under this ordering, the set of policies Π becomes a partially ordered set. A policy is best-effort if it is a maximal element in this partially ordered set, *i.e.*, there is no other policy that dominates it over all $P \in \mathcal{P}$.

To apply Zorn's Lemma (Halmos 1974) and establish the existence of a maximal element, we show that every chain (*i.e.*, a totally ordered subset) $C \subseteq \Pi$ has an upper bound. As mentioned, each policy $\pi \in C$ induces an expected return of $\rho_P^\pi = \langle s_I, V_P^\pi \rangle$, which is a smooth rational function of $P \in \mathcal{P}$. Because the policies in C are totally ordered, their corresponding returns ρ_P^π form a *pointwise increasing* chain: for any $\pi, \pi' \in C$, either $\rho_P^\pi \geq \rho_P^{\pi'}$ or vice versa for all $P \in \mathcal{P}$.

Due to the smoothness and pointwise comparability of these functions, and the compactness of the uncertainty set

---

[3] For details on such solution functions for parametric MDPs, we refer to Junges et al. (2024).

$\mathcal{P}$, we can define the pointwise supremum $\rho_P^{\bar{\pi}}$ of the chain as the return value of a policy $\bar{\pi} \in \mathrm{argmax}_{\pi \in C} \rho_P^\pi$. The policy space Π is the set of all probability distributions over the finite set of actions A and is, hence, compact. Moreover, since the mapping $\pi \mapsto \rho_P^\pi$ is continuous for each fixed P, it follows that the policy $\bar{\pi}$ realizing this supremum exists in Π, *i.e.*, $\bar{\pi} \in \Pi$. That is, $\bar{\pi}$ realizes the pointwise supremum and thus serves as an upper bound of the chain C in Π. Thus, every chain has an upper bound.

Because the policy space is nonempty and every chain has an upper bound, Zorn's Lemma (Halmos 1974) guarantees the existence of at least one maximal element with respect to dominance. By Def. 6, these maximal elements are precisely the best-effort policies. Hence, Π_{BE} is nonempty. $\square$

Theorem 1 (Existence of ORBE policies). *For any RMDP, the intersection of the sets of* optimal robust policies $\Pi^\star$ *and* best-effort policies Π_{BE} *is nonempty.*

Proof. By definition, an optimal robust policy $\pi^\star \in \Pi^\star$, which always exists by construction, maximizes the expected return under the worst-case (fully adversarial) transition function. This means no other policy can strictly dominate $\pi^\star$ in that adversarial transition function, that is, $\nexists \pi \in \Pi \setminus \{\pi^\star\}$ s.t. $\rho_{\mathcal{P}}^\pi > \rho_{\mathcal{P}}^{\pi^\star}$ where $\rho_{\mathcal{P}}^\pi$ and $\rho_{\mathcal{P}}^{\pi^\star}$ represents the robust expected return (as defined in Eq. (4)) for policy π and $\pi^\star$, respectively.

Moreover, when this other policy π is non-optimal robust, *i.e.*, $\pi \in \Pi \setminus \Pi^\star$, we can also deduce that $\pi^\star$ cannot be strictly dominated by π. More precisely, since $\rho_{\mathcal{P}}^{\pi^\star} = \min_{P \in \mathcal{P}} \rho_P^\pi$, it holds that $\rho_{P'}^{\pi^\star} \geq \rho_{\mathcal{P}}^{\pi^\star}$ for all $P' \in \mathcal{P}$. By letting $P' \in \mathrm{argmin}_{P \in \mathcal{P}} \rho_P^\pi$, we thus obtain $\rho_{P'}^{\pi^\star} \geq \rho_{\mathcal{P}}^{\pi^\star} > \rho_{\mathcal{P}}^\pi = \rho_{P'}^\pi$, which implies that $\pi \not\succ_{\mathcal{P}} \pi^\star$.

Then, two cases arise:

1. For all policies $\pi \in \Pi$, $\rho_{\mathcal{P}}^\pi < \rho_{\mathcal{P}}^{\pi^\star}$. In this case, $\pi^\star$ is the unique optimal robust policy and is trivially best-effort as it is a maximal element under strict dominance, and thus cannot be strictly dominated by any other policy (as detailed above). Hence, $\pi^\star \in \Pi_{\mathrm{BE}}$ and so $\Pi^\star \cap \Pi_{\mathrm{BE}} \neq \emptyset$.

2. There exists at least one other policy $\pi' \in \Pi^\star \setminus \{\pi^\star\}$ such that $\rho_{\mathcal{P}}^{\pi'} = \rho_{\mathcal{P}}^{\pi^\star}$. Here, the set of optimal robust policies is not a singleton. Among these policies, there must exist at least one best-effort policy. This follows from the same reasoning based on partial orders and maximal elements as in the proof of Theorem 4. In particular, if no best-effort policy existed, the set of policies ordered by $\geq_{\mathcal{P}}$ would have no maximal element, contradicting the existence of such elements ensured by Zorn's Lemma. Thus, the intersection is nonempty.

In either case, there exists at least one policy that is both optimal robust and best-effort. $\square$

Fig. 4 summarizes this, showing that the orange region (optimal robust policies), the blue region (best-effort policies), and their overlapping area (ORBE policies) are all nonempty.
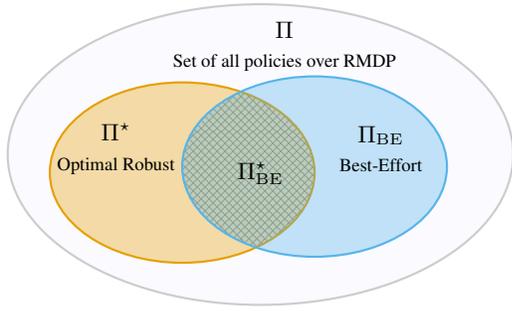
Figure 4: Structure of the policy space in an RMDP. The gray ellipse represents the set of all policies admissible in the RMDP. The orange region denotes the set of *optimal robust* ($\Pi^\star$), while the blue region indicates the set of *best-effort policies* ($\Pi_{\text{BE}}^\star$). The area where the two regions overlap corresponds to the ORBE policies ($\Pi^\star \cap \Pi_{\text{BE}} = \Pi_{\text{BE}}^\star$).

A.2 Proofs of Corollaries 1 to 3

For completeness, we provide the proofs of Corollaries 1 to 3, presented in Sect. 5.

Corollary 1. *Let* $\Pi^\star = \operatorname{argmax}_{\pi \in \Pi} \rho_{\mathcal{P}}^\pi$ *be the set of optimal robust policies. If* $\Pi^\star$ *is a singleton, then* $\pi^\star \in \Pi^\star$ *is ORBE.*

Proof. The proof follows immediately from Theorem 1, which states that the set of optimal robust policies $\Pi^\star$ always contains a best-effort policy. Thus, if there is a single optimal robust policy, this policy must also be best-effort. □

Corollary 2. *Let* $\check{\Pi}^\star = \operatorname{argmax}_{\pi \in \Pi} \rho_{\mathcal{P}}^\pi$ *and let* $\hat{\Pi}^\star = \operatorname{argmax}_{\pi \in \check{\Pi}^\star} \max_{P \in \mathcal{P}} \rho_P^\pi$ *be the set of policies that (within* $\check{\Pi}^\star$*) maximize the expected return under the maximizing* $P \in \mathcal{P}$*. If* $\hat{\Pi}^\star$ *is a singleton, then* $\pi^\star \in \hat{\Pi}^\star$ *is ORBE.*

Proof. First, Theorem 1 states that there exists an ORBE policy within $\check{\Pi}^\star$. Second, the uniqueness of $\pi^\star \in \hat{\Pi}^\star$ implies there exists $P \in \mathcal{P}$ such that $\rho_P^{\pi^\star} > \rho_P^{\pi'}$ for all $\pi' \in \check{\Pi}^\star$. Thus, there is no policy in $\check{\Pi}^\star$ that strictly dominates $\pi^\star$, which proves that $\pi^\star$ is ORBE. □

Corollary 3. *Let* $\bar{\pi} \in \Pi^\star = \operatorname{argmax}_{\pi \in \Pi} \rho_{\mathcal{P}}^{\bar{\pi}}$ *be an optimal robust policy with minimizer* $P^\star \in \operatorname{argmin}_{P \in \mathcal{P}} \rho_P^{\bar{\pi}}$*. Define* $\Pi_{(1)}^\star = \operatorname{argmax}_{\pi \in \Pi^\star} \rho_{P^\star}^\pi$ *and pick a policy* $\pi^\star \in \Pi_{(1)}^\star$*. The policy* $\pi^\star$ *is ORBE if, for all states* $\bar{s}$*, there exists a vector* $\mathbf{v} \in \mathbb{R}^{|S|}$ *such that* $\exists \epsilon > 0$*,* $P_{\bar{s}}^\star + \epsilon \mathbf{v} \in \mathcal{P}_{\bar{s}}$ *and*

$$\nabla_{\mathbf{v}} Z_{P^\star,\bar{s}}^{\pi^\star}(P_{\bar{s}}^\star) > \nabla_{\mathbf{v}} Z_{P^\star,\bar{s}}^{\pi'}(P_{\bar{s}}^\star) \ \forall \pi' \in \Pi_{(1)}^\star \setminus \{\pi^\star\}. \quad (6)$$

Proof. We will show that, for the policy $\pi^\star$, there exists a $P \in \mathcal{P}$ such that $\rho_P^{\pi^\star} > \rho_P^{\pi'}$ for all $\pi' \in \Pi^\star \setminus \{\pi^\star\}$, and thus, $\pi^\star$ is ORBE by Theorem 2.

First, suppose that $\pi' \in \Pi^\star \setminus (\Pi_{(1)}^\star \cup \{\pi^\star\})$. As π' is not in $\Pi_{(1)}^\star = \operatorname{argmax}_{\pi \in \Pi^\star} \rho_{P^\star}^\pi$, it holds that $\rho_{P^\star}^{\pi^\star} > \rho_{P^\star}^{\pi'}$. Thus, $\pi^\star$ cannot be strictly dominated by π'.

On the other hand, suppose that $\pi' \in \Pi_{(1)}^\star \setminus \{\pi^\star\}$. In this case, it holds that $\rho_{P^\star}^{\pi^\star} = \rho_{P^\star}^{\pi'}$. Because the expected return is a smooth function and $\mathbf{v}$ points inside the uncertainty set, the

condition $\nabla_{\mathbf{v}} Z_{P^\star,\bar{s}}^{\pi^\star}(P_{\bar{s}}^\star) > \nabla_{\mathbf{v}} Z_{P^\star,\bar{s}}^{\pi'}(P_{\bar{s}}^\star)$ implies that there exists $\lambda > 0$ such that

$$Z_{P^\star,\bar{s}}^{\pi^\star}(P_{\bar{s}}^\star + \lambda \mathbf{v}) > Z_{P^\star,\bar{s}}^{\pi'}(P_{\bar{s}}^\star + \lambda \mathbf{v}).$$

As this condition holds for every state $\bar{s} \in S$ and the policies $\pi^\star$ and π' must differ in at least one state, it follows that, also in the second case, π' cannot strictly dominate $\pi^\star$. Thus, $\pi^\star$ cannot be strictly dominated by any policy $\pi' \in \Pi^\star \setminus \{\pi^\star\}$, so we conclude that $\pi^\star$ is ORBE by Theorem 2. □

A.3 Rational form of the value function.

Toward the proof of Theorem 3, we show in the following Lemma 1 that the parametric transition function $Z_{P,\bar{s}}^\pi(P_{\bar{s}})$ in Def. 8 is a rational function of degree one (*i.e.*, a fraction of two linear functions). In the proof Lemma 1, we use the *Sherman-Morrison formula*, a well-known matrix identity (Golub and Loan 2013, Sect. 2.1.4), which states that, for a nonsingular matrix $A \in \mathbb{R}^{n \times n}$, vectors $u, v \in \mathbb{R}^n$, and $1 + v^\top A^{-1} u \neq 0$, it holds that

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1} uv^\top A^{-1}}{1 + v^\top A^{-1} u}. \quad (9)$$

In other words, the inverse of a rank-one update to a nonsingular matrix A can be expressed in terms of the inverse of A itself, as long as $1 + v^\top A^{-1} u \neq 0$. For a derivation of Eq. (9), we refer to Golub and Loan (2013, Sect. 2.1.4).

Recall that $P_{\bar{s}} \colon A \to \Delta_S$ is a function from actions to distributions over states. For notational simplicity, we will also interpret $P_{\bar{s}}(a)$, $a \in A$, as a vector in $\mathbb{R}^{|S|}$, and $P_{\bar{s}}$ as a matrix in $\mathbb{R}^{|A| \times |S|}$.

Lemma 1 (Value function as rational). *For any policy* π*, state* $\bar{s} \in S$*, and partial transition function* $P_{-\bar{s}}$*, the value* $Z_{P,\bar{s}}^\pi(P_{\bar{s}})$ *in state* $\bar{s}$ *as a function of the completion* $P_{\bar{s}} \in \mathcal{P}_{\bar{s}}$ *can be written as a rational function of the form*

$$Z_{P,\bar{s}}^\pi(P_{\bar{s}}) = \frac{R^\pi(\bar{s}) + \sum_{a \in A} \alpha_a^\top P_{\bar{s}}(a)}{1 - \sum_{a \in A} \varphi_a^\top P_{\bar{s}}(a)}, \quad (10)$$

where R^π *is defined by Eq. (3), and for all* $a \in A$*, the coefficients* $\alpha_a \in \mathbb{R}_{\geq 0}^{|S|}$ *and* $\varphi_a \in \mathbb{R}_{\geq 0}^{|S|}$ *are defined appropriately.*

Proof. The expected return $V_P^\pi \in \mathbb{R}^{|S|}$ under the policy π and transition function P is written in matrix form as

$$V_P^\pi = (I - \gamma P^\pi)^{-1} R^\pi, \quad (11)$$

where $P^\pi \in \mathbb{R}^{|S| \times |S|}$ and $R^\pi \in \mathbb{R}^{|S|}$ are the matrix and vector forms of Eqs. (2) and (3) over all states, respectively. Define $e_s \in \mathbb{R}^{|S|}$ as the vector with value 1 in entry s only and 0 otherwise, and conversely, define $e_{\neg s} \in \mathbb{R}^{|S|}$ as the vector with value 0 in entry s only and 1 otherwise. Then, Eq. (11) can be decomposed as

$$V_P^\pi = (I - \gamma \cdot \operatorname{diag}(e_{\neg \bar{s}}) P^\pi - \gamma \cdot \operatorname{diag}(e_{\bar{s}}) P^\pi)^{-1} R^\pi. \quad (12)$$

For Lemma 1, we are given a partial transition function $P_{-\bar{s}}$, which thus fixes $P^\pi(s)$ is fixed for all $s \in S \setminus \{\bar{s}\}$. Hence,

by interpreting the completion $P_{\bar{s}}$ as a vector in $\mathbb{R}^{|S|}$, we can rewrite Eq. (12) as

$$V_P^\pi = \left(G - \gamma \cdot e_{\bar{s}} P_{\bar{s}}^\top\right)^{-1} R^\pi, \tag{13}$$

where $G = I - \gamma \cdot \operatorname{diag}(e_{\neg \bar{s}}) P^\pi$ is fixed. As $\gamma < 1$, the matrix G is nonsingular, so we can apply the Sherman-Morrison formula from Eq. (9) (with $A := G$, $u := -\gamma \cdot e_{\bar{s}}$ and $v := P_{\bar{s}}$) to obtain

$$\left(G - \gamma \cdot e_{\bar{s}} P_{\bar{s}}^\top\right)^{-1} = G^{-1} + \frac{G^{-1}(\gamma \cdot e_{\bar{s}} P_{\bar{s}}^\top) G^{-1}}{1 - P_{\bar{s}}^\top G^{-1} \gamma \cdot e_{\bar{s}}},$$

where $P_{\bar{s}}^\top G^{-1} \gamma \cdot e_{\bar{s}} \in [0, 1)$ for $\gamma < 1$ (which we consider; see Def. 1). Thus, V_P^π in Eq. (13) is rewritten as

$$V_P^\pi = \left(G^{-1} + \frac{G^{-1}(\gamma \cdot e_{\bar{s}} P_{\bar{s}}^\top) G^{-1}}{1 - P_{\bar{s}}^\top G^{-1} \gamma \cdot e_{\bar{s}}}\right) R^\pi \tag{14}$$

$$= \left(\frac{G^{-1}\left[1 - P_{\bar{s}}^\top G^{-1} \gamma \cdot e_{\bar{s}} + (\gamma \cdot e_{\bar{s}} P_{\bar{s}}^\top) G^{-1}\right]}{1 - P_{\bar{s}}^\top G^{-1} \gamma \cdot e_{\bar{s}}}\right) R^\pi,$$

which has a numerator and a denominator that are both linear in $P_{\bar{s}}$. Thus, Eq. (14) can be written as a rational function between two linear functions. Let us write $V_P^\pi = [V_P^\pi(s_1), \ldots, V_P^\pi(s_{|S|})]$. By Def. 8, we have that $Z_{P,\bar{s}}^\pi(P_{\bar{s}}) = V_{P_{-\bar{s}} \times P_{\bar{s}}}^\pi(\bar{s}) = V_P^\pi(\bar{s})$, yielding the rational form in Eq. (10) with appropriate coefficients α_a and φ_a for all $a \in Act$. Finally, the domains of α_a and φ_a follow from the fact that the rewards and transition probabilities are nonnegative. $\qquad \square$

Intuitively, the numerator $R^\pi(\bar{s}) + \sum_{a \in A} \alpha_a^\top P_{\bar{s}}(a)$ in Eq. (10) is the sum of the immediate reward and the future discounted reward along paths that do *not* loop back to state $\bar{s}$. Moreover, the term $\sum_{a \in A} \varphi_a^\top P_{\bar{s}}(a)$ in the denominator is the discounted probability of (eventually) looping back to state $\bar{s}$ (and thus acts as a normalizing constant). For any $\gamma < 1$, this probability is strictly smaller than one.

A.4 Proof of Theorem 3
In this section, we use Lemma 1 to provide the proof of Theorem 3, which states that, for any RMDP, our characterization yields an ORBE policy.

Equivalence of policies. First, we show that if two policies attain the same expected return and derivatives under two distinct transition functions, then these policies attain the same expected return on an entire line segment in the space of transition functions. This intuition is formalized by Lemma 2.

Lemma 2 (Equivalence along $P_{\bar{s}}$ line segment). *Let $\pi, \pi' \in \Pi$ be two policies, let $\bar{s} \in S$ be a state, let $P \in \mathcal{P}$ be a transition function, and let $P_{\bar{s}}^{(i)} \in \mathcal{P}_{\bar{s}}$, $i = 1, 2$ be two distinct transition functions in state $\bar{s}$. Define $\mathbf{v} = P_{\bar{s}}^{(2)} - P_{\bar{s}}^{(1)}$. If, for all $i = 1, 2$, it holds that*

$$Z_{P,\bar{s}}^\pi(P_{\bar{s}}^{(i)}) = Z_{P,\bar{s}}^{\pi'}(P_{\bar{s}}^{(i)}), \tag{15a}$$

$$\nabla_{\mathbf{v}} Z_{P,\bar{s}}^\pi(P_{\bar{s}}^{(i)}) = \nabla_{\mathbf{v}} Z_{P,\bar{s}}^{\pi'}(P_{\bar{s}}^{(i)}), \tag{15b}$$

then the expected returns are the same on the entire line segment between points $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$, i.e., for all $\lambda \in [0, 1]$,

$$Z_{P,\bar{s}}^\pi(q) = Z_{P,\bar{s}}^{\pi'}(q) \quad \forall q = \lambda P_{\bar{s}}^{(1)} + (1 - \lambda) P_{\bar{s}}^{(2)}. \tag{16}$$

Proof. The conditions above state that there exist two policies π and π' that attain the same values in two different points $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$. Furthermore, the directional derivatives $\nabla_{\mathbf{v}} Z_{P,\bar{s}}^\pi(P_{\bar{s}}^{(i)})$ and $\nabla_{\mathbf{v}} Z_{P,\bar{s}}^{\pi'}(P_{\bar{s}}^{(i)})$, $i = 1, 2$, in the direction $\mathbf{v}$ of the line segment connecting $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$ are also equal. We will show that these four constraints are sufficient for Eq. (16) to hold.

Recall from Lemma 1 that for any policy π, the value function $Z_{P,\bar{s}}^\pi(P_{\bar{s}})$ in state $\bar{s}$ is a rational function of the form

$$Z_{P,\bar{s}}^\pi(P_{\bar{s}}) = \frac{R^\pi(\bar{s}) + \sum_{a \in A} \alpha_a^\top P_{\bar{s}}(a)}{1 - \sum_{a \in A} \varphi_a^\top P_{\bar{s}}(a)},$$

with appropriate coefficients α_a and φ_a, $a \in Act$. For the proof of Lemma 2, we only need to consider the values of $Z_{P,\bar{s}}^\pi(P_{\bar{s}})$ for values $P_{\bar{s}} \in \{\lambda P_{\bar{s}}^{(1)} + (1 - \lambda) P_{\bar{s}}^{(2)} : \lambda \in [0, 1]\}$. In other words, we restrict the value function to the line segment between $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$. Thus, we may further simplify the multivariable function $Z_{P,\bar{s}}^\pi(P_{\bar{s}})$ as the univariate function $Y_{P,\bar{s}}^\pi \colon [0, 1] \to \mathbb{R}$ defined for all $\lambda \in [0, 1]$ as

$$Y_{P,\bar{s}}^\pi(\lambda) = Z_{P,\bar{s}}^\pi(\lambda P_{\bar{s}}^{(1)} + (1 - \lambda) P_{\bar{s}}^{(2)}) = \frac{\tilde{a} + \tilde{b}\lambda}{\tilde{c} + \tilde{d}\lambda}, \tag{17}$$

with appropriate coefficients $\tilde{a} \in \mathbb{R}, \tilde{b} \in \mathbb{R}, \tilde{c} \in \mathbb{R}, \tilde{d} \in \mathbb{R}$.

Thus, the value function on the line segment between $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$ is defined by four parameters. At the same time, we have four constraints on $Y_{P,\bar{s}}^\pi$, given by

$$Y_{P,\bar{s}}^\pi(0) = \mu_1, \qquad \frac{\partial Y_{P,\bar{s}}^\pi(0)}{\partial \lambda} = \delta_1,$$

$$Y_{P,\bar{s}}^\pi(1) = \mu_2, \qquad \frac{\partial Y_{P,\bar{s}}^\pi(1)}{\partial \lambda} = \delta_2,$$

which together fully define these four coefficients of $Y_{P,\bar{s}}^\pi$.[4] Therefore, any two policies π and π' satisfying the conditions in Eq. (15) must lead to the same coefficients $\tilde{a} \in \mathbb{R}, \tilde{b} \in \mathbb{R}, \tilde{c} \in \mathbb{R}, \tilde{d} \in \mathbb{R}$, and thus satisfy Eq. (16). $\qquad \square$

We will use Lemma 2 to investigate the geometry of the value functions $Z_{P,\bar{s}}^\pi$ and $Z_{P,\bar{s}}^{\pi'}$. For convenience, we simplify the rational function $Z_{P,\bar{s}}^\pi(P_{\bar{s}})$ defined by Eq. (10) as

$$f_{P,\bar{s}}^\pi(x) = \frac{\mathbf{a} + \mathbf{b}^\top x}{1 - \mathbf{c}^\top x}, \tag{18}$$

where $x \in \mathbb{R}^{|A| \cdot |S|}$ represents $P_{\bar{s}}(a)$ concatenated for all $a \in A$, and with the appropriate coefficients $\mathbf{a} \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^{|A| \cdot |S|}$, and $\mathbf{c} \in \mathbb{R}^{|A| \cdot |S|}$. Using this notation, let $L_{P,\bar{s}}^{\pi,\pi'}(x)$ denote the difference between the rationals of two policies π and π, *i.e.*,

$$L_{P,\bar{s}}^{\pi,\pi'}(x) = \frac{\mathbf{a} + \mathbf{b}^\top x}{1 - \mathbf{c}^\top x} - \frac{\mathbf{a}' + (\mathbf{b}')^\top x}{1 - (\mathbf{c}')^\top x} \tag{19}$$

$$= \frac{x^\top \operatorname{diag}(\mathbf{b}'\mathbf{c} - \mathbf{b}\mathbf{c}')x + (\mathbf{a}'\mathbf{c} - \mathbf{a}\mathbf{c}' + \mathbf{b} - \mathbf{b}')^\top x + \mathbf{a} - \mathbf{a}'}{(1 - \mathbf{c}^\top x)(1 - (\mathbf{c}')^\top x)},$$

where the vector-vector multiplication is element-wise.

---

[4]In fact, the function $Y_{P,\bar{s}}^\pi$ is invariant to joint scaling of the parameters and is, thus, already uniquely defined by three constraints.

Proof of Theorem 3. Observe that Eq. (19) is a rational function in x of degree one in the numerator and two in the denominator. Under the conditions required in Lemma 2, the function $L_{P,\bar{s}}^{\pi,\pi'}(x)$ is zero on a line segment. In other words, the numerator in Eq. (19) is zero on a particular line segment. Crucially, observe that a quadratic function is zero on a line segment *only if the quadratic term cancels out, i.e.,* if $\mathbf{b}'\mathbf{c} - \mathbf{b}\mathbf{c}' = 0$. This fact leads to the following theorem, which is our final characterization of ORBE policies.

Theorem 3 (Computing ORBE policies). *Let $\Pi^{\star} = \operatorname{argmax}_{\pi \in \Pi} \rho_{\mathcal{P}}^{\pi}$ be the optimal robust policies. Pick two transition functions $P^{(1)}, P^{(2)} \in \mathcal{P}$ such that $P^{(1)} \neq P^{(2)}$ and, for all $\bar{s} \in S$, the line $g_{\bar{s}}(\lambda) = \lambda P_{\bar{s}}^{(1)} + (1-\lambda)P_{\bar{s}}^{(2)}$ intersects the relative interior[5] of $\mathcal{P}_{\bar{s}}$, or $\mathcal{P}_{\bar{s}} \cap \{g_{\bar{s}}(\lambda)\} = \mathcal{P}_{\bar{s}}$. Define*

$$\Pi_{(1)}^{\star} = \operatorname*{argmax}_{\pi \in \Pi^{\star}} \rho_{P^{(1)}}^{\pi}, \qquad \Pi_{(2)}^{\star} = \operatorname*{argmax}_{\pi \in \Pi_{(1)}^{\star}} \rho_{P^{(2)}}^{\pi}.$$

Choose a policy $\pi^{\star} \in \Pi_{(2)}^{\star}$ s.t., for all $\bar{s} \in S$, it holds that

$$\nabla_{\mathbf{v}} Z_{P^{(1)},\bar{s}}^{\pi^{\star}}(P_{\bar{s}}^{(1)}) \geq \nabla_{\mathbf{v}} Z_{P^{(1)},\bar{s}}^{\pi'}(P_{\bar{s}}^{(1)}) \ \forall \pi' \in \Pi_{(2)}^{\star}, \quad (7a)$$

$$\nabla_{\mathbf{v}} Z_{P^{(2)},\bar{s}}^{\pi^{\star}}(P_{\bar{s}}^{(2)}) \leq \nabla_{\mathbf{v}} Z_{P^{(2)},\bar{s}}^{\pi'}(P_{\bar{s}}^{(2)}) \ \forall \pi' \in \Pi_{(2)}^{\star}, \quad (7b)$$

where the vector $\mathbf{v} \in \mathbb{R}^{|S|}$ is defined as

$$\mathbf{v} = \begin{cases} P^{(2)} - P^{(1)} & \text{if } \rho_{P^{(2)}}^{\pi^{\star}} > \rho_{P^{(1)}}^{\pi^{\star}}, \\ P^{(1)} - P^{(2)} & \text{otherwise.} \end{cases}$$

Then, the policy $\pi^{\star}$ is ORBE

Proof. Observe that the policy $\pi^{\star}$ satisfies one of the following three points:

1. $\hat{\Pi}^{\star}$, $\hat{\Pi}_{(1)}^{\star}$, or $\hat{\Pi}_{(2)}^{\star}$ is a singleton, so that $\pi^{\star}$ is ORBE by Corollary 1 or Corollary 2;
2. $\hat{\Pi}^{\star}$, $\hat{\Pi}_{(1)}^{\star}$, and $\hat{\Pi}_{(2)}^{\star}$ are no singletons but either Eq. (7a) or Eq. (7b) holds with strict inequality, so that $\pi^{\star}$ is ORBE by Corollary 3;
3. $\hat{\Pi}^{\star}$, $\hat{\Pi}_{(1)}^{\star}$, and $\hat{\Pi}_{(2)}^{\star}$ are no singletons and Eq. (7) with non-strict inequality.

We will prove Theorem 3 by showing that, even in the third case, the policy $\pi^{\star}$ is ORBE. As $\Pi_{(2)}^{\star}$ is not unique, there exists a policy $\pi' \in \Pi_{(2)}^{\star} \setminus \{\pi^{\star}\}$ that, for all $\bar{s} \in S$, satisfies:

$$Z_{P^{(1)},\bar{s}}^{\pi^{\star}}(P_{\bar{s}}^{(1)}) = Z_{P^{(1)},\bar{s}}^{\pi'}(P_{\bar{s}}^{(1)}),$$

$$Z_{P^{(2)},\bar{s}}^{\pi^{\star}}(P_{\bar{s}}^{(2)}) = Z_{P^{(2)},\bar{s}}^{\pi'}(P_{\bar{s}}^{(2)}),$$

$$\nabla_{\mathbf{v}} Z_{P^{(1)},\bar{s}}^{\pi^{\star}}(P_{\bar{s}}^{(1)}) = \nabla_{\mathbf{v}} Z_{P^{(1)},\bar{s}}^{\pi'}(P_{\bar{s}}^{(1)})$$

$$\nabla_{\mathbf{v}} Z_{P^{(2)},\bar{s}}^{\pi^{\star}}(P_{\bar{s}}^{(2)}) = \nabla_{\mathbf{v}} Z_{P^{(2)},\bar{s}}^{\pi'}(P_{\bar{s}}^{(2)}).$$

Hence, observe that the policies $\pi^{\star}$ and π' satisfy the conditions in Lemma 2. In addition, Theorem 3 requires that, for all $\bar{s} \in S$, the line $g_{\bar{s}}(\lambda)$ between $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$ either intersects the relative interior of $\mathcal{P}_{\bar{s}}$, or completely covers $\mathcal{P}_{\bar{s}}$.

[5]The relative interior of a convex set X is defined as $\operatorname{relint}(X) \coloneqq \{x \in X : \forall y \in X, \exists \lambda > 1. \ \lambda x + (1-\lambda)y \in X\}$.

First, consider the case where $g_{\bar{s}}(\lambda)$ covers $\mathcal{P}_{\bar{s}}$, *i.e.,* $\mathcal{P}_{\bar{s}} \cap \{g_{\bar{s}}(\lambda)\} = \mathcal{P}_{\bar{s}}$. In this case, Lemma 2 implies that the policies $\pi^{\star}$ and π' have the same value in the entire uncertainty set. As a result, it holds that $\rho_P^{\pi'} = \rho_P^{\pi^{\star}}$ for all $P \in \mathcal{P}$, so the policy $\pi^{\star}$ is ORBE.

Second, consider the case where the line $g_{\bar{s}}(\lambda)$ intersects the relative interior of $\mathcal{P}_{\bar{s}}$, that is, there exists $\lambda \in [0,1]$ such that $g_{\bar{s}}(\lambda) \in \operatorname{relint}(\mathcal{P}_{\bar{s}})$. We use the definition of $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x)$ in Eq. (19) to prove this case by contradiction. As such, suppose that the other policy π' strictly dominates $\pi^{\star}$. In this case, there must exist a $P' \in \mathcal{P}_{\bar{s}}$ where $\rho_{P'}^{\pi'} > \rho_{P'}^{\pi^{\star}}$. By Lemma 2, the point P' cannot be on the line segment between $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$. Because the line $g_{\bar{s}}(\lambda)$ between $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$ intersects the relative interior of $\mathcal{P}_{\bar{s}}$, there also exists another point $P'' \in \mathcal{P}$ such that the line through P' and P'' is perpendicular to $g_{\bar{s}}(\lambda)$. Moreover, as the quadratic term of $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x)$ is zero (*i.e.,* it is a rational of degree one) and $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x)$ has a value of zero on $g_{\bar{s}}(\lambda)$, the fact that $\rho_{P'}^{\pi'} > \rho_{P'}^{\pi^{\star}}$ implies that $\rho_{P''}^{\pi'} < \rho_{P''}^{\pi^{\star}}$. In other words, the existence of a point P' where π' has *higher* expected return than $\pi^{\star}$, implies the existence of another point P'' where π' has *lower* expected return than $\pi^{\star}$. Therefore, such a policy π' that strictly dominates $\pi^{\star}$ cannot exist, so $\pi^{\star}$ is ORBE. □

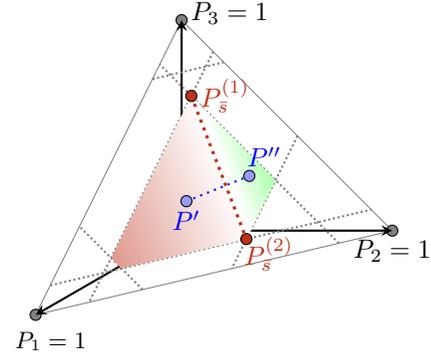A visualization of the proof of Theorem 3 is given in Fig. 5.

Figure 5: Visualization of the proof of Theorem 3 for a convex polytopic uncertainty set $\mathcal{P}_{\bar{s}}$ over three states. The line segment between $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$ is shown in red. The color shade in the polytope depicts the difference $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x)$ in value between the policies $\pi^{\star}$ and π'. Red means $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x) < 0$, white means $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x) = 0$, and green means $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x) > 0$. Because $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x)$ is zero along the line segment and intersects the interior of the uncertainty set $\mathcal{P}_{\bar{s}}$, for every point P' where $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x) < 0$ (*i.e.,* π' outperforms $\pi^{\star}$), there exists another point P'' where $L_{P,\bar{s}}^{\pi^{\star},\pi'}(x) > 0$ (*i.e.,* π' performs worse than $\pi^{\star}$). In particular, this point $P'' \in \mathcal{P}_{\bar{s}}$ can be chosen to be any point such that the line through P' and P'' is perpendicular to the line through $P_{\bar{s}}^{(1)}$ and $P_{\bar{s}}^{(2)}$.

B Details on Empirical Evaluation

In this appendix, we provide further details about the models used in the empirical evaluation and the implementation of robust value iteration that we use.

B.1 Gridworld Models

We generate *slippery gridworlds* of different sizes and with different numbers of obstacles, such as the instance shown in Fig. 6. The objective for the agent is to minimize the expected number of steps to reach the target (in green) from the initial state (in blue). Upon hitting an obstacle (in red), the agent resets to the initial state.

Interval MDP. For the interval MDP (IMDP) used in Sect. 6, we use the model depicted in Fig. 7. For every direction (left, right, up, down), the agent can choose between two actions: one where the slipping probability p is *fixed*, and one where it belongs to the *interval* $[q, p]$. This model structure is repeated for every cell in the grid.

s-Rectangular RMDP. For the RMDP used in Sect. 6.2, we use the model depicted in Fig. 8. Similar to the example RMDP in Fig. 1, this gridworld RMDP has an s-rectangular uncertainty set, which is, in this case, parametrized by the maximum slipping probability p and an improvement q. The value of p is fixed, *e.g.*, $p = 0.25$, whereas the value of q belongs to an interval, *e.g.*, $0 \leq q \leq 0.25$. Thus, for $q = 0$, both action types yield the same value, whereas the best-effort action dominates the non-best-effort action for any $q > 0$.

B.2 Robust Value Iteration

In our experiments, we use two implementations of robust value iteration: one for IMDPs within the probabilistic model checker PRISM (Kwiatkowska, Norman, and Parker 2011), and one for s-rectangular RMDPs that we implemented ourselves in Python. Our own implementation of robust value iteration follows the standard form also described in Sect. 2. That is, given an initial policy $\pi \in \Pi$ and uncertainty set $P \in \mathcal{P}$, we iterate between the following steps:

1. Given fixed P, for every state $s \in S$, update the policy $\pi(s)$ by maximizing the value $V(s)$ in state s:

$$\pi(s) \leftarrow \underset{\pi(s) \in \Delta_A}{\operatorname{argmax}} \left\{ R^\pi(s) + \langle \gamma P^\pi(s), V \rangle \right\},$$

$$V(s) \leftarrow \max_{\pi(s) \in \Delta_A} \left\{ R^\pi(s) + \langle \gamma P^\pi(s), V \rangle \right\}.$$

2. Given fixed π and V, for every state $s \in S$, update the worst-case transition function:[6]

$$P(s, \cdot) \leftarrow \underset{P(s, \cdot) \in \mathcal{P}_s}{\operatorname{argmin}} \left\{ R^\pi(s) + \langle \gamma P^\pi(s), V \rangle \right\},$$

which we compute by solving a linear optimization program (under the assumption that $\mathcal{P}_s$ is a convex polytope).

For the gridworld experiments in Sect. 6, the goal is to compute a policy that *minimizes* the expected return. Thus, for these experiments, we replace each max with min in the algorithm above and vice versa.

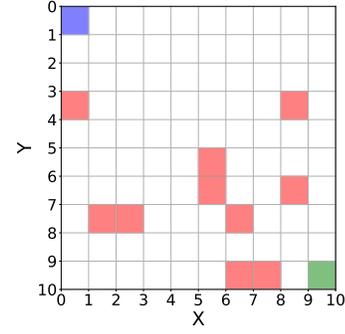[6]Here, we use $P(s, \cdot)$ to denote the transition probabilities in state $s \in S$ for all actions $a \in A$.

Figure 6: Example instance of the slippery gridworld model of size 10×10 and with 10 obstacles (in red), initial state (in blue), and target (in green). Upon hitting a target, the agent is reset to the initial state.
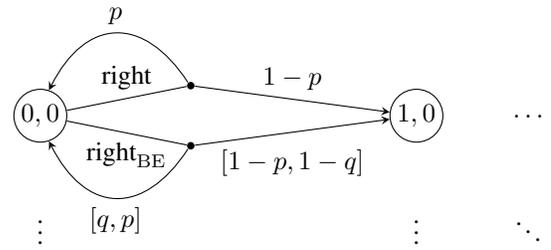

Figure 7: Structure of a single action in the slippery gridworld IMDP model used in Sect. 6. For every move into one of the four cardinal directions (left, right, up, down), the agent has a *normal* action with a fixed slipping probability p (*e.g.*, right), and a *best-effort* action (*e.g.*, right$_{\mathrm{BE}}$) with a slipping probability interval $[q, p]$. The structure shown is repeated for every cell in the grid.
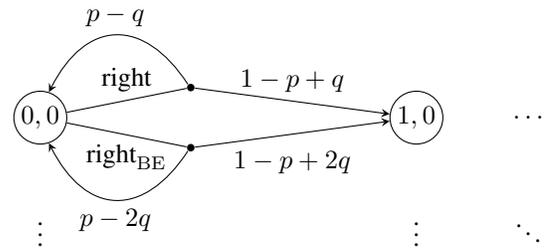

Figure 8: Structure of a single action in the slippery gridworld RMDP model used in Sect. 6.2. This model has an s-rectangular uncertainty set, defined by a fixed worst-case slipping probability (*e.g.*, $p = 0.25$) and an uncertain improvement in the slipping probability of, *e.g.*, $0 \leq q \leq 0.1$. The structure shown is repeated for every cell in the grid.

As discussed in Sect. 2, randomized policies are necessary to achieve optimal values in s-rectangular RMDPs in general. However, for the gridworld models, deterministic policies are sufficient for optimality, meaning that we may replace the (arg)max over distributions $\pi(s) \in \Delta_A$ in step 1 by an (arg)max over actions $a \in A$.

Parameters.   In our implementation, we alternate between these steps until the value function has converged (up to a predefined $\epsilon > 0$, where we use $\epsilon = 10^{-4}$ in our experiments), or a predefined number of iterations is reached (we use a limit of $1\,000$ in our experiments). In practice, this leads to a non-optimized but functional implementation of robust value iteration, which we use to compute policies and evaluate our methods for s-rectangular RMDPs.

B.3   Results

As described in Sect. 6, we repeat every instance over 10 random seeds. Note that the only source of randomness in our experiments comes from the generation of obstacles and the randomization in the order of defining the best-effort versus non-best-effort actions (reflected by the parameter ν). All standard deviations of the results presented in Table 1 and 2 are negligible and are thus omitted for clarity.