Solving Long-run Average Reward Robust MDPs via Stochastic Games

Krishnendu Chatterjee\textsuperscript{1}, Ehsan Kafshdar Goharshady\textsuperscript{1}, Mehrdad Karrabi\textsuperscript{1}, Petr Novotn\textsuperscript{y}\textsuperscript{2}, Đorđe Žikelić\textsuperscript{3}*  

\textsuperscript{1}Institute of Science and Technology Austria (ISTA), Austria  
\textsuperscript{2}Masaryk University, Czech Republic  
\textsuperscript{3}Singapore Management University, Singapore  
{krishnendu.chatterjee, ehsan.goharshady, mehrdad.karrabi}@ist.ac.at, petr.novotny@fi.muni.cz, dzikelic@smu.edu.sg

Abstract

Markov decision processes (MDPs) provide a standard framework for sequential decision making under uncertainty. However, MDPs do not take uncertainty in transition probabilities into account. Robust Markov decision processes (RMDPs) address this shortcoming of MDPs by assigning to each transition an uncertainty set rather than a single probability value. In this work, we consider polytopic RMDPs in which all uncertainty sets are polytopes and study the problem of solving long-run average reward polytopic RMDPs. We present a novel perspective on this problem and show that it can be reduced to solving long-run average reward turn-based stochastic games with finite state and action spaces. This reduction allows us to derive several important consequences that were hitherto not known to hold for polytopic RMDPs. First, we derive new computational complexity bounds for solving long-run average reward polytopic RMDPs, showing for the first time that the threshold decision problem for them is in NP \(\cap\) \text{coNP} and that they admit a randomized algorithm with sub-exponential expected runtime. Second, we present Robust Polytopic Policy Iteration (RPPI), a novel policy iteration algorithm for solving long-run average reward polytopic RMDPs. Our experimental evaluation shows that RPPI is much more efficient in solving long-run average reward polytopic RMDPs compared to state-of-the-art methods based on value iteration.

1 Introduction

Robust Markov decision processes. Markov decision processes (MDPs) are widely adopted as a framework for sequential decision making in environments that exhibit stochastic uncertainty [Puterman, 1994]. The goal of solving an MDP is to compute a policy which maximizes expected payoff with respect to some reward objective. However, classical algorithms for solving MDPs assume that transition probabilities are known and available. While this is plausible from a theoretical perspective, in practice this assumption is not always justified – transition probabilities are typically inferred from data and as such their estimates come with a certain level of uncertainty. This has motivated the study of robust Markov decision processes (RMDPs) [Nilim and Ghaoui, 2003; Iyengar, 2005]. RMDPs differ from MDPs in that they only assume knowledge of some uncertainty set containing true transition probabilities. The goal of solving RMDPs is to compute a policy which maximizes the worst-case expected payoff with respect to all possible choices of transition probabilities belonging to the uncertainty set. The advantage of considering RMDPs rather than MDPs for planning and sequential decision making tasks is that solving RMDPs provides policies that are robust to potential uncertainty in the underlying environment model.

Long-run average reward RMDPs. Recent years have seen significant interest in solving RMDPs. Two most fundamental reward objectives defined over unbounded or infinite time horizon trajectories are the discounted-sum reward and the long-run average reward. While there are many works on solving discounted-sum reward RMDPs (see an overview of related work below), the problem of solving long-run average reward RMDPs remains comparatively less explored. However, long-run average reward is, along with discounted-sum, one of the two fundamental objectives used in agent planning and reinforcement learning [Puterman, 1994; Sutton et al., 1999]. Unlike discounted-sum, which places more emphasis on earlier rewards and prefixes of a run, long run average reward focuses on long-term sustainability of the reward signal.

Prior work and challenges. Most prior work on solving long-run average reward RMDPs focuses on special instances of RMDPs in which uncertainty sets are either products of intervals (i.e. \(L^\infty\)-balls) or \(L^1\)-balls around some pivot probability distribution. These works typically focus on studying Blackwell optimality in RMDPs. In standard MDPs, Blackwell optimality [Puterman, 1994] is a classical result which says that there exists some threshold discount factor \(\gamma^* \in [0, 1)\) such that the same policy maximizes expected discounted-sum reward for any discount factor \(\gamma \in [\gamma^*, 1)\) while also maximizing the long-run average reward in the MDP. This allows using the methods for discounted-sum reward MPDs to solve long-run average reward MDPs.
The work of [Tewari and Bartlett, 2007] considers RMDPs in which uncertainty sets are products of intervals (also called interval MDPs [Givan et al., 2000]), and is the first to establish Blackwell optimality for interval MDPs. It also proposes a value iteration algorithm for solving long-run average reward interval MDPs. The works [Grand-Clément and Petrik, 2023; Goyal and Grand-Clément, 2023] establish Blackwell optimality for RMDPs with general polytopic uncertainty sets. Moreover, the work [Grand-Clément and Petrik, 2023] designs value and policy iteration algorithms for solving long-run average reward RMDPs with interval or $L^1$-uncertainty sets. Finally, the recent work of [Wang et al., 2023] establishes Blackwell optimality and proposes a value iteration algorithm for a general class of RMDPs in which uncertainty sets are only assumed to be compact, but in which the RMDP is assumed to be unichain.

While these works present significant advances in solving long-run average reward RMDPs, our understanding of this problem still suffers from a few limitations:

- **Computational complexity.** While prior works focus on studying Blackwell optimality in RMDPs and algorithms for solving them, no upper bound on the computational complexity of solving this problem is known.
- **Structural or uncertainty restrictions.** The value iteration algorithm of [Wang et al., 2023] requires the RMDP to be unichain. On the other hand, other algorithms discussed above are applicable to RMDPs with either interval or $L^1$-uncertainty sets.
- **Policy iteration.** In classical MDP and stochastic games [Puterman, 1994; Hansen et al., 2013] and in discounted-sum MDPs [Sinha and Ghatre, 2016; Ho et al., 2021], the number of iterations of policy iteration is known to be more efficient than value iteration. However, for RMDPs with uncertainty sets that are not intervals or $L^1$-balls, the only available algorithm is that of [Wang et al., 2023] which is based on value iteration.

**Our contributions.** This work focuses on computational complexity aspects and efficient algorithms for solving long-run average reward RMDPs and our goal is to address the limitations discussed above. To that end, we consider polytopic RMDPs in which all uncertainty sets are assumed to be polytopes and which unify and strictly subsume RMDPs with interval or $L^1$-uncertainty sets. We present a new perspective on the problem of solving long-run average reward polytopic RMDPs by showing that it admits a linear-time reduction to the problem of solving long-run average reward turn-based stochastic games (TBSGs) in which both state and action sets are finite. The reduction is linear in the size of the RMDP and the number of vertices of uncertainty polytopes. While the connection between RMDPs and TBSGs has been informally mentioned in several prior works, e.g. [Nilim and Ghaoui, 2003; Tessler et al., 2019; Goyal and Grand-Clément, 2023], we are not aware that any prior work has actually formalized this reduction. The reduction formalization turns out to be technically non-trivial and we provide a first formal proof of the correctness of the reduction. The reduction formalization is the main technical contribution of our work.

While definitely compelling, the reduction is not only of theoretical interest. TBSGs have been widely studied within computational game theory and formal methods communities [Shapley, 1953; Filar and Vrieze, 1996; Kucera, 2011] and this reduction allows us to leverage a plethora of results on TBSGs in order to establish new results on long-run average reward RMDPs that were hitherto not known to be true. In particular, our results allow us to overcome all the limitations discussed above. Our main contributions are as follows:

1. We establish novel complexity results on solving long-run average reward polytopic RMDPs, showing for the first time that the threshold decision problem for them is in $\text{NP} \cap \text{coNP}$ and that they admit a randomized algorithm with sub-exponential expected runtime. Moreover, for RMDPs with constant number of actions and treewidth of the underlying graph, the problem admits an algorithm with quasi-polynomial runtime.

2. By utilizing Blackwell optimality for TBSGs and the fact that policy iteration has already been shown to be sound for discounted-sum reward TBSGs, we propose Robust Polytopic Policy Iteration (RPPI), our novel algorithm for solving long-run average reward polytopic RMDPs. Our algorithm does not impose structural restrictions on the RMDP such as unichain or aperiodic.

3. **Experimental evaluation.** We implement RPPI and experimentally compare it against the value iteration algorithm of [Wang et al., 2023]. We show that RPPI achieves significant efficiency gains on unichain RMDPs while also being practically applicable to RMDPs without such structural restrictions.

**Related work.** The related work on solving long-run average reward RMDPs was discussed above, hence we omit repetition. In what follows, we overview a larger body of existing works on solving discounted-sum reward RMDPs. Existing approaches can be classified into model-based and model-free. Model-based approaches assume the full knowledge of the uncertainty set and compute optimal policies with exact values on the worst-case expected payoff that they achieve, see e.g. [Nilim and Ghaoui, 2003; Iyengar, 2005; Wiesemann et al., 2013; Kaufman and Schaefer, 2013; Tamar et al., 2014; Ghavamzadeh et al., 2016; Ho et al., 2021]. These works also provide deep theoretical insight into RMDPs and show that many classical results on MDPs transfer to the RMDP setting. Model-free approaches, on the other hand, do not assume the full knowledge of the uncertainty set. Rather, they learn the uncertainty sets from data and as such provide more scalable and practical algorithms, although without guarantees on optimality of learned policies [Roy et al., 2017; Tessler et al., 2019; Wang and Zou, 2021; Wang and Zou, 2022; Ho et al., 2018; Yang et al., 2022].

In the final days of preparing this submission, we found a concurrent work of [Grand-Clément et al., 2023] which also explores the connection between long-run average reward RMDPs and TBSGs in order to study Blackwell optimality. Our focus in this work is on computational complexity aspects and efficient algorithms for solving this problem. Furthermore, we also establish that our reduction is efficient, along with its formal correctness, that allows transfer of algorithms from the stochastic games literature to RMDPs.
2 Preliminaries and Models

For a set $X$, we denote by $\Delta(X)$ the set of all probability distributions over $X$.

MDPs. A (finite-state, discrete-time) Markov decision process (MDP) [Puterman, 1994] is a tuple $(S, A, \delta, r, s_0)$, where $S$ is a finite set of states, $A$ is a finite set of actions, $\delta: S \times A \rightarrow \Delta(S)$ is a probabilistic transition function, $r: S \times A \rightarrow \mathbb{R}$ is a reward function, and $s_0 \in S$ is the initial state. We abbreviate the transition probability $\delta(s, a)(s')$ by $\delta(s' | s, a)$.

The dynamics of an MDP are defined through the notion of a policy. A policy is a map $\sigma : (S \times A)^* \times S \rightarrow \Delta(A)$ which to each finite history of state-action pairs ending in a state assigns the probability distribution over the actions. Under a policy $\sigma$, the MDP evolves as follows. The MDP starts in the initial state $s_0$. Then, in any time step $t \in \{0, 1, 2, \ldots \}$, based on the current history $h_t = s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t$, the policy $\sigma$ samples an action $a_t$ to play according to the probability distribution $\delta(h_t)$. The agent then receives the immediate reward $r_t = r(s_t, a_t)$ and the MDP transitions into the next state $s_{t+1}$ sampled according to the probability distribution $\delta(s_{t+1} | s_t, a_t)$. The process continues in this manner ad infinitum.

A policy is said to be:
- positional (or memoryless), if the choice of the probability distribution over actions depends only on the last state in the history, i.e. if $\sigma(s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t) = \sigma(s_0, a_0, \ldots, s'_{t-1}, a'_{t-1}, s'_t)$ whenever $s_t = s'_t$;
- pure, if for every history $h$ the probability distribution $\sigma(h)$ assigns probability 1 to a single action in $A$.

Robust MDPs. Robust MDPs generalize MDPs in that they prescribe to each state-action pair an uncertainty set, which contains a set of probability distributions over the states in $S$. Thus, MDPs present a special case of robust MDPs in which the uncertainty set prescribed to each state-action pair contains only a single probability distribution over the states.

Formally, a robust MDP (RMDP) [Nilim and Ghaoui, 2003; Iyengar, 2005] is a tuple $(S, A, P, r, s_0)$, where $S$ is a finite set of states, $A$ is a finite set of actions, $P \subseteq \Delta(S) \times A$ is an uncertainty set, $r: S \times A \rightarrow \mathbb{R}$ is a reward function, and $s_0 \in S$ is the initial state.

The intuition behind RMDPs is that, at each time step, the environment policy $\pi$ prescribe to each finite history of state-action pairs an element of the uncertainty set $\mathcal{P}$. Under the agent policy $\sigma$ and the environment policy $\pi$, the RMDP evolves analogously as an MDP defined above, with the difference that in each time step $t \in \{0, 1, 2, \ldots \}$, the state $s_{t+1}$ is sampled from $\delta_t(s_t, a_t)$, where $\delta_t = \pi(s_0, a_0, \ldots, a_{t-1}, s_t, a_t)$.

We note that our definition differs a bit from those found in the literature, where the environment policies have limited adaptiveness: they are either stationary [Ho et al., 2021] or Markovian (i.e., time-dependent) [Wang et al., 2023]. However, one of the corollaries of our results is that all three definitions lead to equivalent problems in the setting we consider.

We denote by $\mathbb{P}_M^\sigma$ and $\mathbb{E}_M^\sigma$ the probability measure and the expectation operator induced by the pair of policies $\sigma, \pi$ over the space of infinite trajectories $s_0, a_0, s_1, a_1, \ldots$ of $\mathcal{M}$.

3 Reduction to Turn-based Stochastic Games

This section presents the main result of this work. We show that this problem of solving long-run average reward polytopic RMDPs admits a linear-time reduction to the problem of solving long-run average reward turn-based stochastic games. The reduction is linear in the size of the RMDP and the number of vertices of uncertainty polytopes. In what follows, we
first recall turn-based stochastic games. We then formally define and prove our reduction, which is the main theoretical result of this work. Finally, we use the reduction to obtain a sequel of novel results on RMDPs by leveraging results on turn-based stochastic games.

3.1 Background on Turn-based Stochastic Games

Turn-based stochastic games are a standard model of decision making in the presence of both adversarial agent and stochastic uncertainty [Shapley, 1953; Filat and Vrieze, 1996; Kučera, 2011]. Formally, a (finite-state, two-player, zero-sum) turn-based stochastic game (TBSG) is a tuple \( G = (S, A, \delta, r, s_0) \) where all the elements are defined as in MDPs, with the additional requirement that the state set \( S \) is partitioned into sets \( S_{\text{max}} \) and \( S_{\text{min}} \) and the action set \( A \) is partitioned into sets \( A_{\text{max}} \) and \( A_{\text{min}} \). We say that these vertices and actions belong to players Max and Min, respectively. Similarly to MDPs and RMDPs, the dynamics of TBSGs are defined through the notion of policies for each player. We distinguish between Max- and Min-histories depending on whether the history ends in a state belonging to Max or Min.

The value of a policy pair, optimal value and optimal policy for long-run average and discounted-sum objectives are defined analogously as for RMDPs. We denote by \( \Sigma \) the sets of all policies of Max and Min, respectively. Similarly, we denote by \( \Sigma_p \) and \( \Pi \), the sets of all pure positional policies of Max and Min, respectively. The following is a formal definition. Let \( \mathcal{M} = (S, A, r, s_0) \) be a polytopic RMDP. We define a TBSG \( G_{\mathcal{M}} = (S_{\text{max}}^{\mathcal{M}}, A_{\text{max}}^{\mathcal{M}}, \delta_{\text{max}}^{\mathcal{M}}, r^{\mathcal{M}}, s_0) \) as below, and say that \( G_{\mathcal{M}} \) is induced by \( \mathcal{M} \). In what follows, for each state-action pair \((s, a) \in S \times A\) in the RMDP, we use \( V_{s,a}^{\mathcal{M}} \) to denote the set of all vertices of the uncertainty polytope \( P_{s,a}^{\mathcal{M}} \): 

- **States.** \( S_{\text{max}}^{\mathcal{M}} \) is defined as a union of two disjoint sets \( S_{\text{max}}^{\mathcal{M}} = S_{\text{max}}^{\mathcal{M}} \cup S_{\text{min}}^{\mathcal{M}} \), where \( S_{\text{max}}^{\mathcal{M}} = S_{\text{max}}^{\mathcal{M}} \) and \( S_{\text{min}}^{\mathcal{M}} = S_{\text{max}}^{\mathcal{M}} \times A_{\text{max}}^{\mathcal{M}} \) are copies of the sets of all states and all state-action pairs in \( \mathcal{M} \).

- **Transition function.** \( \delta_{\text{max}}^{\mathcal{M}} \) is defined as a union of two disjoint sets \( \delta_{\text{max}}^{\mathcal{M}} = \cup_{(s,a) \in S \times A_{\text{max}}^{\mathcal{M}}} V_{s,a}^{\mathcal{M}} \) are copies of the sets of all actions and all vertices of uncertainty polytopes in \( \mathcal{M} \).

- **Reward function.** \( s^{\mathcal{M}} : S_{\text{max}}^{\mathcal{M}} \times A_{\text{max}}^{\mathcal{M}} \cup S_{\text{min}}^{\mathcal{M}} \times A_{\text{min}}^{\mathcal{M}} \rightarrow \mathbb{R} \) is defined as follows:
achieve the optimal value for each player in the TBSG, we can Since by Theorem 1 there exist pure positional policies that

one-to-one correspondences when restricted to pure positional policies, these mappings hold for each $\sigma$ icy pairs under long-run average objective, i.e. that

\[ \text{Val}_G^* \left( \sigma^* \right) = s_0^* \in s_{\text{max}}, \text{i.e. the initial state in } \mathcal{G}_M \text{ is the copy of the initial state in } M. \]

The following theorem establishes reduction correctness. In the proof, considering polytopic RDFMs allows us to soundly prune the action space of the environment up to corners of the polytope. This is then used to reduce the RDFM to a TBSG with finite state and action space.

**Theorem 2** (Correctness). Consider a polytopic RDFM $M = (S^M, A^M, P^M, r^M, s^0_M)$ and define a TBSG $G_M = (S_G^M, A_G^M, \delta_G^M, r_G^M, s^0_G^M)$ as above. Then, for long-run average objective we have $\text{Val}^*(M) = \text{Val}^*(G_M)$.

Proof sketch, full proof in [Chatterjee et al., 2023]. Denote by $\Sigma_M$ and $\Pi_M$ the sets of all the policies and the environment in the polytopic RDFM, and similarly define $\Sigma_G^M$ and $\Pi_G^M$ in the TBSG. We use $\Sigma_p$, $\Pi_p$, $\Sigma_p^G$ and $\Pi_p^G$ to denote the subsets of pure positional policies. In the full proof, we first define mappings from the sets of agent and environment policies in the polytopic RDFM to the sets of agent player and environment player policies in the TBSG:

\[ \Phi : \Sigma_M \to \Sigma_G^M, \quad \Psi : \Pi_M \to \Pi_G^M, \]

we then show that these mappings preserve the values of policy pairs under long-run average objective, i.e. that

\[ \text{Val}(M, \sigma, \pi) = \text{Val}(G_M, \Phi(\sigma), \Psi(\pi)) \]

hold for each $\sigma \in \Sigma_M$ and $\pi \in \Pi_M$. Finally, we show that when restricted to pure positional policies, these mappings give rise to one-to-one correspondences (i.e. bijections)

\[ \Phi_p : \Sigma_p \to \Sigma_p^G, \quad \Psi_p : \Pi_p \to \Pi_p^G. \]

Since by Theorem 1 there exist pure positional policies that achieve the optimal value for each player in the TBSG, we can conclude that there exist pure positional policies that achieve the optimal value in the polytopic RDFM and that the optimal values in the polytopic RDFM and the TBSG coincide.

In order to study the computational complexity of the reduction, we first need to fix a representation of polytopic RDFMs. In particular, we assume that each polytopic RDFM is represented as an ordered tuple $M = (S^M, A^M, \bar{P}^M, r^M, s^0_M)$, where $S^M$ is a list of states, $A^M$ is a list of actions, $\bar{P}^M$ is a list of polytope vertices for each state-action pair, $r^M$ is a list of rewards for each state-action pair and $s^0_M$ is an element of the state list. The following theorem shows that the above is a linear-time reduction in the size of the RDFM and the number of vertices of uncertainty polytopes.

**Theorem 3** (Complexity, proof in [Chatterjee et al., 2023]). Consider a polytopic RDFM $M = (S^M, A^M, \bar{P}^M, r^M, s^0_M)$ and define a TBSG $G_M = (S_G^M, A_G^M, \delta_G^M, r_G^M, s^0_G^M)$ as above. Then, the size of $G_M$ is linear in the size of $M$ and the number of vertices of uncertainty polytopes.

### 3.3 Discussion and Implications

Our reduction enables us to leverage the rich repository of results about TBSGs achieved within the computational game theory and formal methods communities. Since our reduction yields a well-defined and effective correspondence between policies in the polytopic RDFM and the TBSG, the results for TBSGs naturally carry over to the polytopic RDFM setting. In this subsection we highlight several such results, focusing on the issues of complexity and efficient algorithms.

**Pure positional determinacy.** First, as an immediate corollary of our reduction and of Theorem 1, we get the following result. While this result was already established for polytopic RDFMs by a direct analysis and without the reduction to TBSGs [Goyal and Grand-Clément, 2023], we state it below as it will be used for proving subsequent results.

**Corollary 1.** In a polytopic RDFM under long-run average objective, both the agent and the environment have pure positional optimal policies.

**Problem complexity.** The decision problem associated with solving long-run average reward TBSGs (such as deciding whether the optimal value of the game is at least a given threshold $v$) are known to be in $\text{NP} \cap \text{coNP}$. This is due to Theorem 1: for NP membership, one can guess an optimal pure positional policy of, say player Max, and then verify that it enforces value $\geq v$; the verification can be done in polynomial time since fixing a pure positional policy in an RDFM yields a standard MDP that can be solved in polynomial time e.g. by linear programming [Puterman, 1994]. The argument for $\text{coNP}$ membership is dual. By Corollary 1, the same characterization holds for RDFMs.

**Efficient algorithms.** Simple stochastic games (SSGs) [Condon, 1992] are a special class of TBSGs with reachability objectives that captures the complexity of the whole TBSG class. It was shown by [Andersson and Milternes, 2009] that solving SSGs with both discounted and long-run average payoffs is polynomial-time reducible to solving SSGs (where by solving we mean computing optimal values and policies).

While SSGs are not known to be solvable in polynomial time, several sub-exponential algorithms for them were developed. The randomized algorithm due to [Ludwig, 1995] solves a given SSG in time $O(2^{\sqrt{\log(|S_{\text{max}}|)}, \log(|G|))$, where $|G|$ is the encoding size of the game. Neither our reduction from RDFMs to TBSGs nor the reduction from TBSGs to SSGs [Andersson and Milternes, 2009] add any player Max vertices. Hence, we get the following result.

**Corollary 2.** There is a randomized algorithm for solving long-run average reward polytopic RDFMs whose expected runtime is $O(2^{\sqrt{|S|} \cdot \log(|M|))}$, i.e., sub-exponential.

**Remark 1** (Discounted polytopic RDFMs). In the extended version of the paper [Chatterjee et al., 2023a], we show the above reduction can be modified to also work for discounted-sum objectives. Hence Corollaries 1–3 hold for discounted-sum objectives as well.
In [Chatterjee et al., 2023b] authors provide an algorithm for solving TBSGs parameterized by the treewidth of the underlying graph. The algorithm has runtime $O((tn^2 \log nW)^{\log t})$ where $n$ is the number of nodes, $t$ is the treewidth of the graph and $W$ is the maximum absolute value of rewards in the game. Let $M$ be an RMDP with $m$ actions and whose underlying graph has treewidth $t$. In the extended version of the paper [Chatterjee et al., 2023a], we show that $G_M$ has treewidth at most $(t+1)(m+1)−1$, which implies the following result.

**Corollary 4.** Given an RMDP $M$ with constant number of actions and treewidth of the underlying graph, the long-run average objective can be solved in quasi-polynomial time.

### 4 Algorithm for Long-run Average RMDPs

In this section, we present our algorithm **Robust Polytopic Policy Iteration (RPPI)** for computing long-run average values and optimal pure positional policies in polytopic RMDPs. The pseudocode of our algorithm is shown in Algorithm 1. Our algorithm can be viewed as a policy iteration algorithm for solving long-run average reward polytopic RMDPs. Policy iteration was hitherto not known to be sound for solving long-run average reward RMDPs, thus the design of a policy iteration algorithm is another contribution of our work.

**Outline.** RPPI uses the reduction in Section 3 to reduce our problem to solving long-run average TBSGs, for which policy iteration is known to be sound and highly efficient [Hansen et al., 2013]. Our RPPI is thus motivated by efficient implementations of policy iteration for TBSGs, which utilize Blackwell optimality and use policy iteration to solve a number of discounted-sum TBSGs while increasing the discount factor until a discount factor for which discounted-sum and long-run average policies are equal is achieved.

**Algorithm details.** Algorithm 1 takes as input a polytopic RMDP $M$ (line 1). It then constructs a TBSG $G_M$ as in Section 3 to reduce our problem to solving a long-run average TBSG (line 2), initializes a discount factor $\gamma = \frac{1}{2}$ (line 3) and performs policy iteration for long-run average TBSGs (lines 4-10). For each value of the discount factor $\gamma$, Algorithm 1 first solves the discounted TBSG and computes a pair of optimal policies by using the discounted-sum TBSG policy iteration algorithm of [Hansen et al., 2013], which runs in polynomial time, as an off-the-shelf subprocedure (line 5). It then uses Policy Profile Evaluation (PPE, see details below) to check if this policy profile is optimal for the LIMAVG objective as well (line 6). If the policy profile is found to be optimal, then the algorithm breaks the loop (lines 7-8) and returns the agent policy together with its LIMAVG value (line 12-13). Otherwise, the discount factor $\gamma$ is increased (line 10) and Algorithm 1 proceeds to solving the next discounted-sum TBSG (line 4-10). By Blackwell optimality for TBSGs, Algorithm 1 is guaranteed to eventually terminate and find a pair of optimal pure positional policies for each player in the TBSG $G_M$. By the one-to-one correspondence between pure positional policies in the RMDP $M$ and the TBSG $G_M$ established in the proof of Theorem 2 and in Corollary 1, the agent player policy then gives rise to an optimal pure positional agent policy in the RMDP $M$, as desired.

### Algorithm 1 Robust Polytopic Policy Iteration (RPPI)

1: **Input** polytopic RMDP $M$
2: $G_M \leftarrow$ TBSG by reduction from $M$ as in Section 3
3: $\gamma \leftarrow \frac{1}{2}$ (initial discount factor)
4: **while** True **do**
5: $\sigma, \pi \leftarrow$ optimal discounted-sum policies in $G_M$ with discount factor $\gamma$, by method of [Hansen et al., 2013]
6: **if** is_optimal, rewards $\leftarrow$ PPE($G_M, \sigma, \pi$)
7: **if** is_optimal **then**
8: break
9: **end if**
10: $\gamma \leftarrow \frac{1+\gamma}{2}$
11: **end while**
12: $Val^*(M) = rewards_{[s_0]}$
13: **Return** $Val^*(M), \sigma$

**Policy profile evaluation.** We now describe the PPE subprocedure used for policy profile evaluation in line 6 of Algorithm 1. PPE takes as input the TBSG $G_M$ and policies $\sigma$ and $\pi$ of Max and Min players. In order to check whether $\sigma$ and $\pi$ are optimal for the long-run average objective, by pure positional determinacy of TBSGs (Theorem 1) it suffices to check if they provide optimal responses to each other.

To check if $\sigma$ and $\pi$ provide optimal responses to each other, PPE first considers an MDP obtained by fixing policy $\pi$ of the Min player and solves the long-run average MDP to compute the optimal long-run average payoff of Max player against $\pi$. Similarly, PPE then considers the MDP obtained by fixing policy $\sigma$ of the Max player and solves the long-run average MDP to compute the optimal long-run average payoff of Min player against $\sigma$. In both cases, the long-run average MDP can be solved via linear programming or policy iteration [Puterman, 1994]. If two MDPs have equal long-run average payoffs, then we conclude that two policies are optimal responses to each other, otherwise they are not.

The following theorem establishes the correctness of our RPPI algorithm (and the PPE subprocedure). Its proof follows from the correctness of our reduction (Theorem 2), Blackwell optimality for TBSGs [Andersson and Miltersen, 2009], and the correctness of all involved subprocedures for which we use off-the-shelf approaches.

**Theorem 4.** Given a polytopic RMDP $M$, Algorithm 1 returns the optimal long-run average value $Val^*(M)$ and an optimal policy for the agent.

### 5 Experimental Results

We implement RPPI (Algorithm 1) and compare it against two state of the art value iteration-based methods for solving long-run average reward RMDPs with uncertainty sets not being intervals or $L^1$-balls, towards demonstrating the significant computational runtime gains provided by a policy iteration-based algorithm. Furthermore, we demonstrate the applicability of our method to non-unichain polytopic RMDPs to which existing algorithms are not applicable. Our implementation is publicly available at https://github.com/mehrdad76/RMDP-LRA.
Implementation details. We implement the policy iteration for discounted-sum TBSSGs of [Hansen et al., 2013] to compute optimal strategies in line 5 in Algorithm 1. Furthermore, we use the probabilistic model checker Storm [Hensel et al., 2022] as an off-the-shelf tool to compute values in long-run average reward MDPs in the Policy Profile Evaluation sub-procedure. Due to numerical precision issues in Python and Storm, in PPE we replace the equality by $|\text{max\_rewards} - \text{min\_rewards}| < 10^{-5}$. All experiments were run in Python 3.9 on a Ubuntu 22.04 machine with an octa-core 2.40 GHz Intel Core i5 CPU, 16 GB RAM.

Baselines. We compare the runtime of our RPPI against two value iteration-based algorithms: Robust Value Iteration (RVI) and Robust Relative Value Iteration (RRVI), both proposed in [Wang et al., 2023]. Similarly to our RPPI, RVI uses Blackwell optimality to compute the long-run average value and optimal policy by solving discounted-sum reward RMDPs and gradually increasing the discount factor. On the other hand, RRVI is based on a value iteration procedure that directly solves the long-run average reward RMDP.

Both RVI and RRVI impose structural restrictions on the RMDP. RVI requires the RMDP to be a unichain. RRVI requires the RMDP to be unichain and aperiodic. In contrast, our RPPI does not impose any of these assumptions.

Moreover, while RVI and RRVI are guaranteed to converge in the limit, they do not provide a stopping criteria for practical implementation. Hence, in our comparison, we run them both until the absolute difference between their computed values and our computed value is at most $10^{-3}$.

Benchmarks. We consider the following benchmarks:

1. Contamination Model. Contamination Model is a conservative model of an RMDP defined over a given MDP where, at each step, the probability of taking the prescribed MDP transition function is $R \in [0, 1]$ whereas with probability $1 - R$ the environment gets to choose any probability distribution over the states. In our evaluation, we consider a Contamination Model taken from [Wang et al., 2023]. For a given natural number $n$, the base MDP contains $n$ states and $n + 10$ actions. The transition function and rewards for each state-action pair are then randomly generated (see [Wang et al., 2023] for details). The uncertainty set $\mathcal{P}_{(s, a)}$ for state-action pair $(s, a)$ is $\mathcal{P}_{(s, a)} = \{(1 - R)\delta_{(s, a)} + R \cdot p | p \in \Delta(S)\}$. The resulting RMDP is polytopic, unichain and aperiodic, hence both baselines are applicable to it.

2. Frozen Lake. This benchmark modifies the Frozen Lake environment in the OpenAI Gym [Towers et al., 2023] to turn it into an RMDP. Consider an $n \times n$ grid with some cells being holes. An agent starts at the top left corner and chooses an action from $\{\text{left, right, up, down}\}$ to move to the next cell in that direction. However, the grid is slippery and the agent might move in a direction perpendicular to the chosen direction. If the agent falls into a hole, they are stuck and remain there with probability 1. The reward increases as the agent gets closer to the bottom right corner of the grid, so the agent should get to and stay around the bottom right corner. We turn this model into a polytopic RMDP by allowing the adversarial environment to perturb the transition probabilities by increasing the probability of moving to one adjacent cell by at most $d = 0.2$ while decreasing the remaining probabilities uniformly. Note that this RMDP is aperiodic and multichain, hence neither RVI nor RRVI baselines are applicable. We use this environment to evaluate our RPPI on multichain polytopic RMDPs to which existing methods are not applicable. In order to allow for a comparison, we also consider a unichain variant of the Frozen Lake environment where the holes are treated as "walls" and cannot be visited. To distinguish between the two variants, we refer to them as Unichain and Multichain Frozen Lake models, respectively.

Results on Contamination Model. We compared our method against both baselines and our results are shown in Figure 1. As indicated by the results, our RPPI is considerably faster than RVI. Given the similar structure of RPPI and RVI with the main difference being the use policy instead of value iteration, our results confirm the significant gains in computational runtime provided by our reduction to TBSSGs which allowed the design of a policy iteration algorithm. On the other hand, we observe that the RRVI baseline is faster than both our method and the RVI baseline. The explanation for this is that RRVI is an algorithm specifically tailored to unichain aperiodic RMDPs, whereas our RPPI imposes no structural restrictions on the RMDP.

Results on Frozen Lake. Table ?? shows our results on the Frozen Lake model. As mentioned before, due to RMDP structural restrictions imposed by the baselines, RVI is only applicable to the Unichain model whereas RRVI is applicable to neither of the models. Our RPPI imposes no structural restrictions on the RMDP and is thus applicable to both models. As we can see from Table ?? our policy iteration-based RPPI is significantly faster than RVI, with an average runtime of 7.8 seconds. In contrast, the RVI method takes more than 2000 seconds on average, and times out on the $10 \times 10$ model. Moreover, our method successfully solves all the Multichain model instances in less than 4 minutes each.

Concluding remarks. We considered the problem of solving long-run average reward RMDPs with polytopic uncertainty sets and proposed a new perspective on this problem by establishing a linear-time reduction to long-run average
reward turn-based stochastic games. This reduction allowed us to leverage results from the stochastic games literature and obtain a number of results on computational complexity and efficient algorithms that were hitherto not known to hold for long-run average polytopic RMDPs. Our work leaves several interesting venues for future work. A natural question to consider is whether and how RMDPs with non-rectangular uncertainty sets could be related to stochastic games. Another direction would be to consider the connection between RMDPs with non-polytopic uncertainty sets and stochastic games, e.g., when uncertainty sets are specified in terms of KL-divergence from some nominal distribution [Wang et al., 2023]. The extension to non-polytopic uncertainty sets via either (a) approximating non-polytopic with polytopic uncertainty sets and applying our method, or (b) designing new approaches, are interesting directions of future work.

Acknowledgements

This work was supported in part by the ERC-2020-CoG 863818 (FoRM-SMArt) and the Czech Science Foundation grant no. GA23-06963S.

References


Table 1: Runtime comparison on the Frozen Lake models.


