Efficient Solution and Learning of Robust Factored MDPs

Yannik Schnitzer, Alessandro Abate, David Parker

University of Oxford
Department of Computer Science
{yannik.schnitzer,alessandro.abate,david.parker}@cs.ox.ac.uk

Abstract

Robust Markov decision processes (r-MDPs) extend MDPs by explicitly modelling epistemic uncertainty about transition dynamics. Learning r-MDPs from interactions with an unknown environment enables the synthesis of robust policies with provable (PAC) guarantees on performance, but this can require a large number of sample interactions. We propose novel methods for solving and learning r-MDPs based on *factored* state-space representations that leverage the independence between model uncertainty across system components. Although policy synthesis for factored r-MDPs leads to hard, non-convex optimisation problems, we show how to reformulate these into tractable linear programs. Building on these, we also propose methods to learn factored model representations directly. Our experimental results show that exploiting factored structure can yield dimensional gains in sample efficiency, producing more effective robust policies with tighter performance guarantees than state-of-the-art methods.

Code — https://zenodo.org/records/17580296
Extended version — https://arxiv.org/abs/2508.00707

1 Introduction

Markov decision processes (MDPs) are the standard modelling framework for sequential decision-making under uncertainty. However, real-world dynamics are often complex and not fully known. In safety-critical settings, it is therefore essential to reason about *epistemic uncertainty*, due to incomplete knowledge of the environment, and to construct *robust* policies that provide provable performance guarantees on the unknown environment they operate in.

Robust Markov decision processes (r-MDPs) (Wiesemann, Kuhn, and Rustem 2013) extend MDPs by not requiring every transition probability to be known precisely but only restricting them to lie in a given *uncertainty set*. These uncertainty sets are typically derived from data, e.g., observed interactions with the unknown system, as in reinforcement learning (RL). Learning for r-MDPs, however, does not optimise for expected performance alone; rather, it enables the synthesis of policies that are robust with respect to the current epistemic uncertainty in the transition dynamics and provides provable *Probably Approximately Correct* (PAC) guarantees on performance with high confidence (Strehl and Littman 2005; Suilen et al. 2022).

Unlike robust RL approaches, that often focus on heuristic or empirical training for difficult scenarios (Morimoto and Atkeson 2002; Pinto et al. 2017), r-MDP learning operates on explicit uncertainty sets learned from data and yields formal anytime guarantees on worst-case performance under the true but unknown transition model.

A practical limitation of r-MDP learning and policy synthesis, however, is that, to achieve high-confidence performance guarantees, the overall confidence level must be distributed across all transition distributions (Strehl and Littman 2005) or individual transition probabilities (Suilen et al. 2022) being learnt. In large-scale environments, this enforces stringent confidence requirements, requiring a high number of samples to construct tight uncertainty sets that yield effective robust policies with meaningful guarantees.

Many real-world domains come with structural knowledge that permits distinct features of the state space to be modelled independently, giving rise to the model of *factored* MDPs (f-MDPs) (Koller and Parr 1999; Boutilier, Dean, and Hanks 1999). RL algorithms have been extended to exploit this factored structure (Kearns and Koller 1999; Guestrin, Patrascu, and Schuurmans 2002; Strehl 2007), often yielding exponential improvements in sample efficiency over learning in the *flat* (non-factored) representation. While these methods come with PAC guarantees, ensuring that a near-optimal policy is learned with high probability in time polynomial in the factored representation, existing work focuses on expected performance and convergence rather than providing provable guarantees on worst-case performance.

In this work, we introduce a *robust factored MDP* framework, which leverages structural independence to construct uncertainty sets for each state factor rather than for a flat model. We show that robust policy synthesis in this setting leads to intractable non-convex optimisation problems, but that for standard uncertainty classes, such as confidence intervals, L_1 balls and general polytopes, these problems admit exact convex reformulations. To address the computational challenges of the resulting, potentially exponential constraint sets, we leverage convex relaxations that preserve tight performance guarantees while enabling efficient solution. We show that our method synthesises more effective robust policies with high-confidence performance guaran-

tees that are substantially tighter than those of prior factored MDP learning approaches. Furthermore, we show that exploiting the factored structure can improve the sample efficiency of robust policy learning by orders of magnitude compared to state-of-the-art methods in flat representations.

2 Problem Formulation

The set of all probability distributions over a finite set Y is denoted by $\Delta(Y) = \{\mu\colon Y \to [0,1] \mid \sum_{y\in Y} \mu(y) = 1\}$. For convenience, we also represent distributions as vectors in the probability simplex, $(p_1, \ldots, p_{|Y|}) \in \Delta_{|Y|}$, where $p_i = \mu(y_i)$ under a fixed ordering of the elements of Y.

2.1 MDPs and Factored MDPs

A *Markov decision process* (or *MDP*) is a tuple $M = (S, A, T, r)$, where S and A are finite sets of states and actions, $T\colon S \times A \to \Delta(S)$ is a transition probability function, and $r\colon S \times A \to \mathbb{R}$ is a reward function. A *policy* is a mapping $\pi\colon (S \times A)^* \times S \to \Delta(A)$ that resolves the nondeterminism by selecting a distribution over actions based on the current state and past interactions. The interaction between a policy and an MDP induces infinite sequences (or *paths*) of the form $s^0a^0s^1a^1\ldots$, where at each step, the next action is drawn from the distribution assigned by the policy, given the current history prefix, and the next state is drawn from the transition distribution $T(\cdot\,|s,a)$.

A *factored MDP* (or *f-MDP*) is an MDP in which states are represented as vectors of n components. Each *factor* i (also referred to as a *state variable* or *state marginal*) takes values from a finite domain $\mathcal{D}_i$. Hence, states are tuples $s = (s_1, \ldots, s_n)$, with $s_i \in \mathcal{D}_i$. To capture the (in-)dependence between factors, we adopt the framework of Strehl (2007). Given an arbitrary set $\mathcal{I}$ of *dependency identifiers*, a function $D_i\colon S \times A \to \mathcal{I}$ is a *dependency function* for factor i. The transition function is then defined as

$$T(s'|s,a) = \prod_{i=1}^{n} P(s'_i|D_i(s,a)), \tag{1}$$

where s'_i denotes the i-th component of the next state s' and each $P(\cdot\,|D_i(s,a)) \in \Delta(\mathcal{D}_i)$ specifies the *marginal probability distribution* of the respective factor.

Example 1. A classic example of a factored MDP is the System Administrator domain (Guestrin, Patrascu, and Schuurmans 2002), where an administrator controls a total of n machines or factors, each of which can be either operational or in a failure state. Each machine is connected to a subset of the others, and its probability of failing at the next step depends on whether its connected neighbours are operational, but is independent of all other machines. The dependency identifiers for machine i thus capture the current state of the machine itself and those of its connected neighbours: if one or more of these neighbours are in a failure state, the marginal probability that machine i fails increases.

2.2 Robust Factored MDPs

Robust factored MDPs (or *rf-MDPs*) (Delgado et al. 2009; Liu, Wiesemann, and Yue 2024) extend factored MDPs to incorporate epistemic uncertainty about transition dynamics. They generalise fixed marginal transition distributions $P(\cdot\,|D_i(s,a)) \in \Delta(\mathcal{D}_i)$ to *marginal uncertainty sets* $\mathcal{P}(D_i(s,a)) \subseteq \Delta(\mathcal{D}_i)$. The overall uncertainty set of possible transition distributions at (s,a) is then defined as:

$$\mathcal{T}(s,a) = \bigotimes_{i=1}^{n} \mathcal{P}(D_i(s,a)), \tag{2}$$

where $\otimes$ denotes the outer product (or Kronecker product) of distributions, extended to sets. Specifically, for sets $\mathcal{P} \subseteq \Delta(\mathcal{D}_i)$ and $\mathcal{Q} \subseteq \Delta(\mathcal{D}_j)$, the product is defined as

$$\mathcal{P} \otimes \mathcal{Q} = \{P \otimes Q \mid P \in \mathcal{P},\, Q \in \mathcal{Q}\}, \tag{3}$$

where for distributions $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_k)$, their outer product is given by

$$(P \otimes Q)_{ij} = p_i \cdot q_j, \quad 1 \le i \le m, 1 \le j \le k. \tag{4}$$

Hence, $\mathcal{T}(s,a)$ comprises all product distributions over the factor-wise uncertainty sets, providing a structured representation of the uncertainty over the full state space S.

As in standard robust MDPs (Nilim and Ghaoui 2005; Iyengar 2005; Wiesemann, Kuhn, and Rustem 2013), rf-MDPs introduce an additional step in the evolution of the process: at a given state s, before the next state is determined following the selection of action a, an *environment policy* τ selects, for each factor i, a marginal distribution from the corresponding uncertainty set $\mathcal{P}(D_i(s,a))$. These combine into a product distribution, as per Eq. (1), which lies in the overall uncertainty set $\mathcal{T}(s,a)$ and defines the probability distribution from which the successor state is drawn.

Objectives. An *objective* is a mapping R that assigns a return to each infinite path $\rho = s^0a^0s^1a^1\ldots$ in an rf-MDP $\tilde{M}$. Given a pair of agent and environment policies π and τ, we denote by $\mathbb{E}^{\pi,\tau}_{\tilde{M},s}$ the induced expectation over paths starting in state s (Wolff, Topcu, and Murray 2012). The *value* of s under π and τ with respect to objective R is defined as

$$V^{\pi,\tau}_{\tilde{M}}(s) = \mathbb{E}^{\pi,\tau}_{\tilde{M},s}[R]. \tag{5}$$

Unless stated otherwise, our results are agnostic to the specific choice of objective. The most common objective is the discounted cumulative reward:

$$R(\rho) = \sum_{t=0}^{\infty} \gamma^t r(s^t, a^t), \tag{6}$$

for some discount factor $0 < \gamma < 1$. However, our results readily extend to other objectives, such as undiscounted rewards (Schwartz 1993; Puterman 1994; Meggendorfer, Weininger, and Wienhöft 2025a) or reachability goals focussing on the probability of eventually reaching a target set of states, possibly whilst avoiding certain undesirable states.

Robust Values and Policies The *optimal robust policy* π^* in an rf-MDP $\tilde{M}$ is the policy that achieves, in every state, the *optimal robust value* $V^*_{\tilde{M}}(s)$, which is the best possible value under the worst-case environment policy. Formally:

$$V^*_{\tilde{M}}(s) = \sup_{\pi} \inf_{\tau} V^{\pi,\tau}_{\tilde{M}}(s), \text{ and} \tag{7}$$

$$\pi^* = \operatorname*{argsup}_{\pi} \inf_{\tau} V^{\pi,\tau}_{\tilde{M}}(s). \tag{8}$$

In this paper, we implicitly assume that the agent aims to maximise the objective while the environment adversarially seeks to minimise it. All results remain valid under the dual case with reversed roles (Nilim and Ghaoui 2005). It is straightforward to verify that the policy π^* guarantees at least the value $V^*_{\tilde{M}}(s)$ on any concrete f-MDP obtained by fixing specific distributions from the uncertainty sets.

Next, in Section 3, we present novel methods for efficiently and accurately solving rf-MDPs, i.e., computing optimal robust values and policies, assuming polytopic marginal uncertainty sets, such as the commonly used L_1, L_∞ balls and general L_p balls. Then, in Section 4, we leverage these methods to efficiently learn robust policies with provable performance guarantees in unknown f-MDPs.

3 Solving Robust Factored MDPs

As for standard robust MDPs, the optimal value function $V^*_{\tilde{M}}$ and a corresponding robust policy for an rf-MDP can be computed with *robust value iteration* (Iyengar 2005; Nilim and Ghaoui 2005). Assuming *rectangular* uncertainty sets, meaning that each state–action pair has an independent uncertainty set over which the environment can act adversarially, the global problem decouples into a local optimisation at every state. For any state s, the agent selects an action $a \in A$ that maximises the worst-case expected return over all transition kernels in $\mathcal{T}(s,a)$, yielding the robust Bellman equation, where $V^*_{\tilde{M}}(s)$ equals:

$$\max_{a \in A} \underbrace{\min_{T \in \mathcal{T}(s,a)} \Big[r(s,a) \; + \gamma \sum_{s' \in S} T(s'|s,a)\, V^*_{\tilde{M}}(s') \Big]}_{\text{Inner Optimisation}}. \quad (9)$$

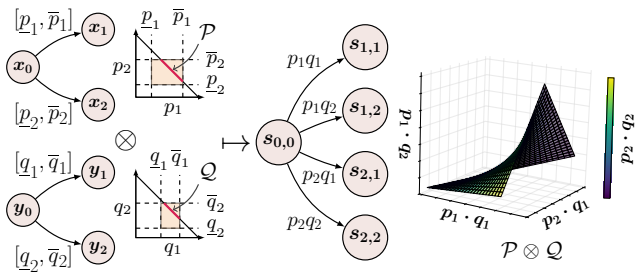
The inner optimisation captures the environment's adversarial choice of a transition kernel within $\mathcal{T}(s,a)$. For standard (non-factored) robust MDPs, this is tractable when $\mathcal{T}(s,a)$ has a favourable geometry: e.g., an L_1 or L_∞ ball, which is solvable via bisection in time linear-logarithmic in the support size (Strehl and Littman 2005), or a polytope described by a number of vertices or half-spaces that can be solved via linear programming (Nilim and Ghaoui 2005).

rf-MDPs, however, induce uncertainty sets $\mathcal{T}(s,a)$ as the multilinear product of marginal sets (see Equation (2)). Even if every marginal $\mathcal{P}(D_i(s,a))$ is convex, convexity is in general not preserved under the product; consequently, $\mathcal{T}(s,a)$ can in general be non-convex (see Figure 1), rendering the inner optimisation hard and often intractable.

We show that when the marginals are polytopes, the associated non-linear problem admits an exact linear reformulation whose constraints follow directly from the polytopic descriptions of the marginals. However, the number of resulting constraints can grow rapidly for many common classes of uncertainty sets. To retain tractability, we construct tight linear overapproximations of $\mathcal{T}(s,a)$, yielding robust Bellman updates that allow for an efficient and accurate solution.

3.1 Exact Products of Polytopic Uncertainty Sets

We consider polytopic marginal uncertainty sets $\mathcal{P}$ defined as the convex hull of finitely many extreme distributions,



(a) Marginal States and Uncertainty Sets

(b) Product States and Uncertainty Set

Figure 1: Part (a) shows two factors of an rf-MDP, with convex marginal uncertainty sets $\mathcal{P}$ and $\mathcal{Q}$, which are line segments in the two-dimensional probability simplex. The resulting product uncertainty set $\mathcal{P} \otimes \mathcal{Q}$ in (b) is non-convex.

i.e., $\mathcal{P} = \text{conv}\{P^{(1)}, \ldots, P^{(m)}\} := \{\sum_{i=1}^m \lambda_i P^{(i)} \,|\, \lambda_i \geq 0,\ \sum_{i=1}^m \lambda_i = 1\}$. We first prove that the resulting inner optimisation problem in (9), taken over the non-convex product uncertainty set $\mathcal{T}(s,a)$, admits an exact linear reformulation. In contrast to prior approaches for solving robust factored MDPs (Delgado, Sanner, and de Barros 2011), this result allows us to avoid the invocation of an expensive and potentially approximate non-linear solver. It builds on two key observations: first, by the bilinearity of the Kronecker product $\otimes$ (Horn and Johnson 1991), the convex hull of $\mathcal{T}(s,a)$ is a polytope whose extreme points are precisely the pairwise products of the extreme distributions of the marginal polytopes (Horst and Tuy 1996), and second, the inner optimisation is linear in the transition probabilities and thus attains its optimum at a vertex of the convex hull.

Theorem 1. *Let $\mathcal{P} = \text{conv}\{P^{(1)}, \ldots, P^{(m)}\} \subseteq \Delta_M$ and $\mathcal{Q} = \text{conv}\{Q^{(1)}, \ldots, Q^{(k)}\} \subseteq \Delta_N$ be polytopic marginal uncertainty sets. Then the corresponding non-linear inner optimisation problem in Equation* (9) *attains its optimum at one of the products of the marginal extreme distributions:*

$$\left\{ P^{(i)} \otimes Q^{(j)} \;\middle|\; 1 \leq i \leq m,\ 1 \leq j \leq k \right\}.$$

The full proofs for all presented results are provided in the extended version. Theorem 1 inductively extends to any number of marginals and offers a direct approach to solving the inner optimisation problem *exactly* by enumerating the product vertices induced by the marginal uncertainty sets of each factor. However, the number of such vertices can grow rapidly, rendering explicit enumeration computationally infeasible, even for standard classes of uncertainty sets arising from statistical estimation (Suilen et al. 2024). For example, when the marginal sets are defined as L_1 or L_∞ balls centred around a nominal distribution, the number of vertices per marginal can grow exponentially in the support size. A detailed construction can be found in the extended version.

3.2 Efficient Solutions through Relaxations

To mitigate the potential intractability of the exact inner optimisation, we use *relaxations*, i.e., overapproximations of

the uncertainty set $\mathcal{T}(s, a)$ that trade exactness for tractability. Since the relaxed set is a superset of the true one, the value returned by the relaxed Bellman operator is a *lower bound* on the exact robust value, guaranteeing that the resulting policy never underperforms against any transition kernel in the original uncertainty set. This sound, worst-case guarantee distinguishes our approach from earlier methods for rf-MDPs, which rely on approximate value-function fitting over a fixed basis (Delgado et al. 2009; Delgado, Sanner, and de Barros 2011; Liu, Wiesemann, and Yue 2024). Such schemes provide no formal bound on the policy's performance and therefore cannot in general provide safety guarantees as required in robust learning.

We aim for *tight* relaxations, admitting as few spurious distributions (i.e., distributions not in the true set) as possible. An overly loose relaxation can lead to a pessimistic bound, and result in an unnecessarily conservative policy.

We first consider marginal uncertainty sets that take the form of *boxes* (or *hyper-rectangles*) intersected with the probability simplex, which are generalisations of L_∞ balls.

These arise naturally when individual transition probabilities are estimated from observed data using confidence intervals (Strehl and Littman 2005; Suilen et al. 2022). A box is defined by lower and upper bounds $\underline{p}, \overline{p} \in [0,1]^N$ on each component of a probability distribution, with $\underline{p}_i \leq \overline{p}_i$ for all $1 \leq i \leq N$, yielding the uncertainty set

$$\mathcal{P}_B = \left\{ (p_1, \ldots, p_N) \in \Delta_N \,\middle|\, \underline{p}_i \leq p_i \leq \overline{p}_i \right\}. \tag{10}$$

Robust MDPs defined in this way are called *interval* or *bounded-parameter MDPs* (Givan, Leach, and Dean 2000).

Interval-Arithmetic Relaxation. A natural relaxation for products of distributions in interval MDPs is to use *interval arithmetic*. In fact, this approach is taken in the modelling language of the PRISM tool (Kwiatkowska, Norman, and Parker 2011), which supports compositional modelling of interval MDPs. Given two box-type uncertainty sets $\mathcal{P}_B \subseteq \Delta_M$ and $\mathcal{Q}_B \subseteq \Delta_N$ with respective bounds $\underline{p}, \overline{p} \in [0,1]^M$ and $\underline{q}, \overline{q} \in [0,1]^N$, the corresponding interval-arithmetic relaxation $\mathcal{R}_{ia} \subseteq \Delta_{M \cdot N}$ is defined as

$$\mathcal{R}_{ia} = \left\{ H \in \Delta_{M \cdot N} \,\middle|\, \underline{p}_i\, \underline{q}_j \leq h_{ij} \leq \overline{p}_i\, \overline{q}_j \right\}. \tag{11}$$

While the interval-arithmetic relaxation is tight with respect to each component individually, it fails to capture dependencies across components and can therefore introduce a large amount of spurious distributions (Hashemi, Hermanns, and Turrini 2016; Mathiesen, Haesaert, and Laurenti 2025). In particular, it admits spurious extreme points, potentially leading to overly conservative solutions in the inner optimisation problem, as we demonstrate in the following example.

Example 2. Consider the two box-type uncertainty sets:

$$\mathcal{P}_B = \{(p, 1-p) \in \Delta_2 \mid p \in [0.2, 0.6]\}, \text{and}$$
$$\mathcal{Q}_B = \{(q, 1-q) \in \Delta_2 \mid q \in [0.1, 0.3]\}.$$

Their interval-arithmetic product relaxation $\mathcal{R}_{ia}$ is:

$$[0.02, 0.18] \times [0.14, 0.54] \times [0.04, 0.24] \times [0.28, 0.72] \cap \Delta_4.$$

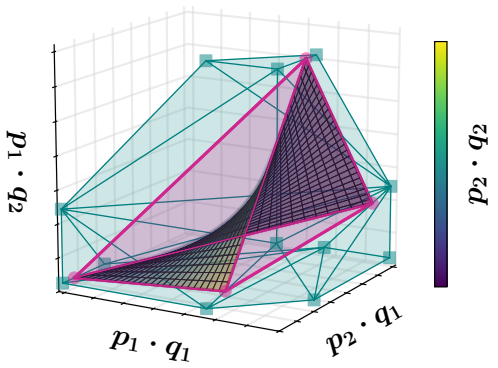


Figure 2: Projections of the interval-arithmetic (blue) and McCormick (pink) relaxations for the product of box-type uncertainty sets (coloured curve). The McCormick relaxation is tighter and has fewer spurious extreme distributions.

Now consider $H = (0.18, 0.14, 0.24, 0.44) \in \mathcal{R}_{ia}$. We can verify that H is a vertex of $\mathcal{R}_{ia}$, as three bounds are tight. Since $h_1 = 0.18 = pq$, the box constraints imply that $p = 0.6$ and $q = 0.3$. But then it must be that:

$$\big(p(1-q),\ (1-p)q,\ (1-p)(1-q)\big) = (0.42,\ 0.12,\ 0.28),$$

so the only valid product distribution with $pq = 0.18$ is $(0.18, 0.42, 0.12, 0.28) \neq H$. Thus H is not contained in the actual product uncertainty set $\mathcal{P}_B \otimes \mathcal{Q}_B$.

McCormick Relaxation. In order to tackle the issue of spurious distributions in interval-arithmetic relaxations, and the conservative solutions to the inner optimisation problem in (9) that may result, we draw on results from non-linear global optimisation and employ *McCormick envelopes* (McCormick 1976). These provide tight convex relaxations of multilinear products through a polynomial number of linear constraints, yielding a tractable linear program that closely approximates the original non-linear formulation.

For two variables $p \in [\underline{p}, \overline{p}]$ and $q \in [\underline{q}, \overline{q}]$, the McCormick envelopes are defined by the following linear inequalities:

$$h \geq p\,\underline{q} + q\,\underline{p} - \underline{p}\,\underline{q}, \tag{12a}$$
$$h \geq p\,\overline{q} + q\,\overline{p} - \overline{p}\,\overline{q}, \tag{12b}$$
$$h \leq p\,\underline{q} + q\,\overline{p} - \overline{p}\,\underline{q}, \tag{12c}$$
$$h \leq p\,\overline{q} + q\,\underline{p} - \underline{p}\,\overline{q}. \tag{12d}$$

Each inequality arises from combining the bounds on p and q. For instance, since $p \geq \underline{p}$ and $q \geq \underline{q}$, we have

$$(p - \underline{p})(q - \underline{q}) \geq 0.$$

Expanding and substituting $h = pq$ gives

$$pq - p\,\underline{q} - q\,\underline{p} + \underline{p}\,\underline{q} \geq 0 \implies h \geq p\,\underline{q} + q\,\underline{p} - \underline{p}\,\underline{q},$$

which is precisely Equation (12a). Despite their simplicity, these inequalities suffice to exactly characterise the convex hull of a single bilinear product $h = pq$ (McCormick 1976).

When applied to the inner optimisation in Equation (9) over a product uncertainty set as per Equation (2), each bilinear term $p_i q_j$ is replaced by an auxiliary variable h_{ij}, which

is constrained by the four McCormick inequalities in (12). We then impose the global simplex constraint $\sum_{i,j} h_{ij} = 1$, ensuring that the auxiliaries $\{h_{ij}\}_{i,j}$ define a valid probability distribution. This reformulation linearises the original non-linear inner optimisation. Figure 2 illustrates how the McCormick relaxation excludes many of the spurious extreme points admitted by the interval-arithmetic relaxation, thus resulting in less conservative solutions and more effective (whilst still robust) policies. Furthermore, since each h_{ij} contributes to exactly four McCormick constraints, the total number of constraints grows only polynomially with the marginal supports, yielding a tractable inner linear program. Full details of this construction and its extension to products of more than two marginal uncertainty sets (obtained by recursive applications) are provided in the extended version.

Relaxations for L_p Uncertainty Sets. The constructions above enable the exact composition of polytopic uncertainty sets and provide tight-yet-tractable relaxations for box-type uncertainty sets. We now also consider uncertainty sets that are L_p norm balls centred at a nominal distribution $\hat{P} \in \Delta_N$, which are typically estimated from observed data as:

$$\mathcal{P}_p(\hat{P}, \varepsilon) = \{P \in \Delta_N \mid \|P - \hat{P}\|_p \leq \varepsilon\}.$$

These sets are generally not polytopic, for $1 < p < \infty$. We hence extend a result from Strehl (2007), originally formulated for the composition of L_1 balls, to arbitrary L_p norms:

Theorem 2. *Let $\mathcal{P}_p(\hat{P}, \varepsilon_1)$ and $\mathcal{P}_p(\hat{Q}, \varepsilon_2)$ be two L_p uncertainty sets for some $1 \leq p \leq \infty$. Then:*

$$\mathcal{P}_p(\hat{P}, \varepsilon_1) \otimes \mathcal{P}_p(\hat{Q}, \varepsilon_2) \subseteq \mathcal{P}_p(\hat{P} \otimes \hat{Q}, \varepsilon_1 + \varepsilon_2).$$

This result offers an approach to solving non-polytopic rf-MDPs, complementing the constructions presented in Section 3.2. When applied to L_1 uncertainty sets, it directly extends the PAC analysis of Strehl (2007) to robust policy synthesis. In Section 5, we compare the various relaxations, showing that our constructions yield substantially tighter uncertainty sets, enable more sample-efficient learning, and deliver robust policies with stronger performance guarantees.

4 Robust Policy Learning in Factored MDPs

We now introduce a novel learning approach that integrates factored model estimation with accurate and tractable robust planning, generating policies that are provably robust for unknown f-MDPs. Based on agent interactions with the environment, we derive marginal uncertainty sets, such as confidence intervals or L_1 balls, which induce a polytopic rf-MDP. Leveraging the solution methods in the previous section, we exploit this factored structure to achieve dimensional gains in sample efficiency compared to existing robust learning methods in flat models, as we demonstrate in our experimental evaluation. Crucially, our approach provides a finite-sample, anytime PAC guarantee: after any number of interactions, we can bound the worst-case performance in the unknown MDP with high confidence.

We consider a factored MDP M with known state space but unknown (marginal) transition distributions. For clarity, we assume that the reward function is known, but all results extend to the case of unknown reward functions (Strehl and Littman 2005). Our algorithm has access to agent-environment interactions in the form of a dataset of transition samples $\mathcal{C} = \{(s_t, a_t, s'_t)\}_t$, where a_t is the action taken in state s_t under some exploration policy and s'_t is the observed successor state. We remain agnostic to the precise sampling mechanism and assume that the sample set $\mathcal{C}$ is given. In Section 5, we describe the specific sampling procedure used in our evaluation.

From the definition of a factored MDP, we first identify the relevant transition components that must be estimated. For a state–action pair (s, a), the relevant dependencies are

$$D_{s,a} = \big\{ j \in \mathcal{I} \;\big|\; \exists i.\ j = D_i(s, a) \big\}.$$

Aggregating over all state–action pairs yields the set of relevant transition components: $\mathcal{Q} = \bigcup_{(s,a) \in S \times A} D_{s,a}$, so that $|\mathcal{Q}|$ counts the number of marginal transition distributions to be estimated. The total number of unknown transition probabilities is the sum of the supports of the marginals: $U = \sum_{j \in \mathcal{Q}} \big|\mathrm{supp}\big(P(\cdot \mid j)\big)\big|$. For a sample dataset $\mathcal{C} = \{(s_t, a_t, s'_t)\}_t$, we define the *realisation counts*:

$$n(x_i, j) = \sum_{(s,a,s') \in \mathcal{C}} \mathbb{1}\big(D_i(s, a) = j \ \wedge\ s'_i = x_i\big),$$

and the *component counts*:

$$n(j) = \sum_{(s,a,s') \in \mathcal{C}} \sum_i \mathbb{1}\big(D_i(s, a) = j\big),$$

for $x_i \in \mathcal{D}_i$ and $j \in \mathcal{I}$. Here, $n(j)$ is the total number of encountered transitions whose transition probability distribution involves a marginal with dependency identifier j, while $n(x_i, j)$ records how often such transitions lead to the marginal state component x_i. From this we can derive the empirical estimates of the marginal distributions as

$$\hat{P}(s'_i | D_i(s, a)) = \frac{n(s'_i, D_i(s, a))}{n(D_i(s, a))}. \tag{13}$$

While this empirical estimate becomes increasingly accurate with more data, it provides no quantification of uncertainty. We aim to synthesise a policy that, after any fixed number of samples, comes with a guaranteed lower bound on its performance in the unknown f-MDP. To achieve this, we inflate each point estimate into a high-confidence uncertainty set over the marginal distribution, thereby defining an rf-MDP.

4.1 Uncertainty Set Construction

We consider two established methods for constructing uncertainty sets. The first builds exact binomial confidence intervals for each transition probability, treating each outcome s'_i under dependency $j = D_i(s, a)$ as a Bernoulli trial (Suilen et al. 2022; Meggendorfer, Weininger, and Wienhöft 2025b). Given $x = n(s'_i, j)$ "successes" in $n = n(j)$ trials and an error probability $\delta \in (0, 1)$, the true transition probability $P(s'_i \mid j)$ lies in the interval:

$$\mathrm{CP}(s'_i, j) = \big[B\big(\tfrac{\delta}{2}; x,\ n - x + 1\big),\ B\big(1 - \tfrac{\delta}{2}; x + 1,\ n - x\big)\big]$$

with probability at least $1-\delta$, where $B(\alpha; u, v)$ denotes the α-quantile of the $\mathrm{Beta}(u, v)$ distribution (Clopper and Pearson 1934). Applying these bounds independently to each transition component defines the box-type uncertainty sets

$$\mathcal{P}(j) = \Big\{ P' \in \Delta(\mathcal{D}_i) \;\Big|\; P'(s'_i) \in \mathrm{CP}(s'_i, j)\ \forall s'_i \Big\},$$

to which our rf-MDP solution techniques apply directly. Throughout, we assume $n(j) > 0$. When $n(j) = 0$, we set the uncertainty sets as the entire probability simplex.

The second approach centres on an L_1-norm ball around the empirical marginal distribution $\hat{P}(\cdot \mid j)$. For each relevant dependency identifier $j = D_i(s, a) \in \mathcal{Q}$, we set

$$\mathcal{P}(j) = \Big\{ P' \in \Delta(\mathcal{D}_i) \;\Big|\; \|P'(\cdot) - \hat{P}(\cdot \mid j)\|_1 \le \varepsilon \Big\},$$

where ε follows from Weissman et al. (2003) as

$$\varepsilon = \sqrt{\frac{2\big[\ln(2^a - 2) - \ln(\delta)\big]}{n(j)}}, \quad a = \big|\mathrm{supp}(P(\cdot \mid j))\big|.$$

This ensures that the true marginal lies in $\mathcal{P}(j)$ with probability at least $1-\delta$. This underpins the native PAC-learning results for both factored and standard MDPs (Strehl and Littman 2005; Strehl 2007). Moreover, it yields polytopic uncertainty sets, as the intersection of an L_1 ball with the probability simplex is a polytope, thus permitting exact composition via Theorem 1. However, L_1 balls do not integrate naturally into the McCormick relaxation without further overapproximating them as boxes. As we show in the extended version, overapproximating L_1 balls by their smallest enclosing box yields a looser uncertainty set than applying the box-type construction directly. Consequently, the radius-sum result of Theorem 2 is the natural choice when composing L_1 marginal sets with a large number of vertices.

4.2 Provably Robust Policy Synthesis

To obtain a provably robust policy with quantifiable performance guarantees in the unknown f-MDP M, we construct an rf-MDP $\tilde{M}$ using the uncertainty sets described above. For the guarantees to be meaningful, we must ensure that the unknown MDP M is *contained* in $\tilde{M}$ (denoted $M \in \tilde{M}$) with high, user-specified confidence. This means that every marginal distribution $P(\cdot \mid j)$ for $j \in \mathcal{Q}$ must lie within its corresponding uncertainty set $\mathcal{P}(j)$.

Given a desired overall confidence probability $1-\beta$, we follow the standard approach of Strehl (2007) and distribute the total error probability $\beta \in (0, 1)$ across all learnt distributions/transitions. Under the L_∞ scheme, this results in $\delta = \beta/U$, and under the L_1 scheme, in $\delta = \beta/|\mathcal{Q}|$. By the union bound, this ensures that $M \in \tilde{M}$ with probability at least $1-\beta$, regardless of the number of observed samples.

When solving the learned rf-MDP $\tilde{M}$ using a *robust*, i.e., either exact or relaxation-based method from Section 3, the following performance guarantee for the resulting robust policy on the true, unknown f-MDP M follows immediately:

Theorem 3. *Let M be an f-MDP and $\tilde{M}$ an rf-MDP such that $\Pr[M \in \tilde{M}] \ge 1-\beta$ for some $\beta > 0$. Let π^* be the policy obtained by solving $\tilde{M}$ with a robust solution method, and let $V^{\pi^*}_{\tilde{M}}(s)$ denote its corresponding robust value. Then,*

$$\Pr\Big[V^{\pi^*}_{M}(s) \;\ge\; V^{\pi^*}_{\tilde{M}}(s)\Big] \;\ge\; 1-\beta. \tag{14}$$

In other words, with probability at least $1-\beta$, the learned robust policy π^* achieves a value in every state of the true f-MDP that is no worse than its computed value in the learned rf-MDP. This PAC-style guarantee based on the novel robust solution methods distinguishes our approach from prior methods (Delgado, Sanner, and de Barros 2011; Liu, Wiesemann, and Yue 2024), which cannot guarantee a valid lower bound, thus forfeiting such a performance guarantee.

5 Experiments

We integrated our methods into the PRISM solver for probabilistic models (Kwiatkowska, Norman, and Parker 2011), which offers a modular language for specifying factored MDPs. We augment PRISM with our algorithms for solving and learning robust factored MDPs and employ the Gurobi optimiser with default parameters for all linear programs.

5.1 Evaluation: Solving rf-MDPs

We evaluate the three methods for *solving* rf-MDPs with box-type uncertainty sets: vertex enumeration, interval-arithmetic relaxations, and McCormick relaxations, across a range of benchmark environments. These include classic f-MDP domains such as the System Administrator domain discussed in Example 1 (Guestrin, Patrascu, and Schuurmans 2002), as well as established r-MDP case studies with inherent factored structure, including multi-agent scenarios like the Aircraft Collision Avoidance domain (Kochenderfer 2015). Detailed descriptions of each domain are provided in the extended version. For each domain, we obtain an rf-MDP by perturbing a nominal transition kernel with an L_∞ uncertainty radius of 0.025 (see the extended version for additional levels of uncertainty), yielding box-type marginal uncertainty sets for each factor.

Results. Table 1 summarises the outcomes. For each method, we report: (i) the robust value of the optimal policy in the rf-MDP; (ii) the runtime to solve the rf-MDP; and (iii) for relaxation-based methods, the relative gap to the exact result obtained by vertex enumeration, quantifying the additional conservatism introduced by over-approximating the product uncertainty sets.

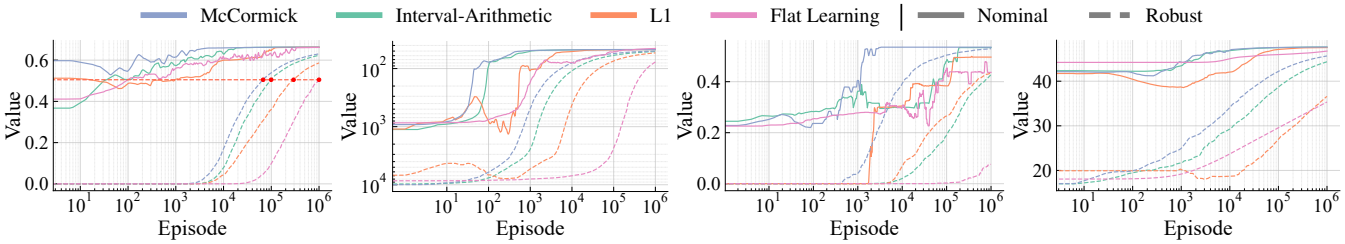
Notably, McCormick relaxations preserve the tightness of vertex enumeration while remaining computationally efficient. Interval-arithmetic relaxations, though generally fast, yield looser bounds due to spurious extreme distributions. Overall, McCormick relaxations strike the best balance between solution tightness and runtime. We present the complete set of experiments, including analyses across varying uncertainty radii in the extended version.

5.2 Evaluation: Robust Policy Learning in f-MDP

We next compare four methods for robust policy *learning*: (i) standard r-MDP learning in the flat model with box-type uncertainty sets; (ii) rf-MDP learning with L_1 uncertainty

Domain	$\lvert S\rvert$	$\lvert T\rvert$	Vertex Enumeration		Interval-Arithmetic			McCormick		
			Robust Value	Time [s]	Robust Value	Rel. Gap	Time [s]	Robust Value	Rel. Gap	Time [s]
Aircraft (↑)	11153	1262099	0.73	2535.8	0.65	11%	6.1	0.73	0%	43.7
Drone (↑)	262144	21694720	0.69	2125.8	0.63	10%	90.2	0.69	0%	190.7
Stock Trading (↑)	12481	5362624	25.43	67.6	17.60	31%	16.0	25.43	0%	67.5
SysAdmin (↑)	15873	9332587	50.70	66.7	46.66	8%	34.1	50.70	0%	64.1
Chain (↓)	100	3136	331.34	778.1	451.28	36%	0.6	331.34	0%	7.6
Frozen Lake (↓)	50625	1866556	216.01	1018.4	242.05	12%	67.7	216.01	0%	105.9
Herman (↓)	2048	177148	20.64	11.0	23.82	15%	2.8	20.64	0%	8.1

Table 1: Results for solving rf-MDPs. Arrows (↑/↓) indicate optimisation directions. $|S|$ and $|T|$ denote the number of states and transitions. The relative gap is $|V_{\text{VE}} - V_R|/V_R$, where V_{VE} and V_R are the robust results from vertex enumeration and respective relaxation. The complete set of experiments, with varying uncertainty radii, can be found in the extended version.



(a) Aircraft (b) Frozen Lake (c) Drone (d) SysAdmin

Figure 3: Results for robust policy learning. The plots show objective value against processed fixed-length trajectories. Dashed curves show the robust guarantee for the learned robust policy, solid curves show its actual performance on the true model. The complete experimental results, including additional domains and total runtimes, are provided in the extended version.

sets solved using the radius-sum result from Theorem 2, which is the direct extension of the PAC analysis of Strehl (2007) to robust policy learning and represents the only available baseline for rf-MDPs; (iii) & (iv) rf-MDP learning with box-type marginal uncertainty sets solved via either interval-arithmetic or McCormick relaxation.

To build the transition dataset $\mathcal{C}$, we iteratively sample fixed-length trajectories that restart in the initial state. To balance exploration and exploitation, we follow the *optimism in the face of uncertainty* principle (Munos 2014), selecting actions that are optimal under the most favourable transition model within the current uncertainty sets. Note that this choice of sampling procedure is arbitrary: the resulting robustness guarantees hold under any alternative sampling strategy, such as random action selection.

Across all domains, we fix the overall confidence level for the inclusion of the true, unknown MDP in the learned r-MDP to $1 - \beta = 0.9999$, (see Equation (14)). Each experiment is repeated with 10 distinct random seeds, and we report the average results along with standard deviation bands.

Results. Figure 3 presents robust policy learning results across various domains. For each method, we plot the robust value of the learned policy (dashed lines) and its nominal performance on the true, hidden model (solid lines) against the number of processed trajectories. While true-model performance provides useful validation, our focus lies on the robust values, i.e., the performance that can be guaranteed with high confidence on the unknown environment.

The results demonstrate significant gains in sample efficiency by exploiting factored structures. Specifically, far fewer fixed-length trajectory samples are required to achieve equivalent robust performance guarantees compared to state-of-the-art methods on flat models. Furthermore, rf-MDP learning with box-type uncertainty sets, derived from exact confidence intervals and solved via convex relaxations, consistently outperforms approaches based on L_1 uncertainty sets and the radius-sum method. McCormick relaxations need about half the number of samples of interval-arithmetic relaxations for the same robust guarantees. This advantage is particularly crucial in domains where data collection is inherently limited, costly, or challenging.

Figure 3a (red line) shows the number of samples needed to match the performance guarantee from flat learning on the Aircraft domain after 10^6 trajectories. Factored learning with L_1 uncertainty sets reduces this to $3 \cdot 10^5$. Interval-arithmetic relaxation further decreases it to 10^5, and McCormick relaxation is the most efficient, requiring only $6 \cdot 10^4$ trajectories. This gap becomes even more pronounced in other domains. We provide the full set of experiments including additional domains, total runtimes and detailed comparisons of sample efficiency in the extended version.

6 Conclusion

We have presented novel methods for solving robust factored MDPs, facilitating exact solutions and optimal robust policies for polytopic uncertainty sets. Utilising global optimisation techniques, we developed relaxation-based approaches that balance accuracy and computational tractability. Our experimental results show that these methods markedly improve accuracy in solving rf-MDPs and enable significantly more sample-efficient robust policy learning.

Acknowledgements

This work was partially supported by the ARIA projects SAINT and SUPER MARTINGALE CERTIFICATES, the UKRI AI Hub on Mathematical Foundations of AI and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 834115, FUN2MODEL). The authors are grateful to Karan Mukhi for the insightful discussions on this work.

References

Boutilier, C.; Dean, T. L.; and Hanks, S. 1999. Decision-Theoretic Planning: Structural Assumptions and Computational Leverage. *J. Artif. Intell. Res.*, 11: 1–94.

Clopper, C. J.; and Pearson, E. S. 1934. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4): 404–413.

Delgado, K. V.; de Barros, L. N.; Cozman, F. G.; and Shirota, R. 2009. Representing and Solving Factored Markov Decision Processes with Imprecise Probabilities. In *Proceedings of the 6th International Symposium on Imprecise Probability: Theories and Applications*, 169–178.

Delgado, K. V.; Sanner, S.; and de Barros, L. N. 2011. Efficient solutions to factored MDPs with imprecise transition probabilities. *Artif. Intell.*, 175(9-10): 1498–1527.

Givan, R.; Leach, S. M.; and Dean, T. L. 2000. Bounded-parameter Markov decision processes. *Artif. Intell.*, 122(1-2): 71–109.

Guestrin, C.; Patrascu, R.; and Schuurmans, D. 2002. Algorithm-Directed Exploration for Model-Based Reinforcement Learning in Factored MDPs. In *ICML*, 235–242. Morgan Kaufmann.

Hashemi, V.; Hermanns, H.; and Turrini, A. 2016. Compositional Reasoning for Interval Markov Decision Processes. *CoRR*, abs/1607.08484.

Horn, R. A.; and Johnson, C. R. 1991. *Topics in Matrix Analysis*. Cambridge University Press.

Horst, R.; and Tuy, H. 1996. *Global Optimization: Deterministic Approaches*. Springer Series in Operations Research. Springer.

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

Kearns, M. J.; and Koller, D. 1999. Efficient Reinforcement Learning in Factored MDPs. In *IJCAI*. Morgan Kaufmann.

Kochenderfer, M. 2015. *Decision Making Under Uncertainty: Theory and Application*.

Koller, D.; and Parr, R. 1999. Computing Factored Value Functions for Policies in Structured MDPs. In *IJCAI*, 1332–1339. Morgan Kaufmann.

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.

Liu, H.; Wiesemann, W.; and Yue, M. 2024. An MILP-Based Solution Scheme for Factored and Robust Factored Markov Decision Processes. *CoRR*, abs/2404.02006.

Mathiesen, F. B.; Haesaert, S.; and Laurenti, L. 2025. Scalable control synthesis for stochastic systems via structural IMDP abstractions. In *HSCC*, 14:1–14:12. ACM.

McCormick, G. P. 1976. Computability of global solutions to factorable nonconvex programs: Part I - Convex underestimating problems. *Math. Program.*, 10(1): 147–175.

Meggendorfer, T.; Weininger, M.; and Wienhöft, P. 2025a. Solving Robust Markov Decision Processes: Generic, Reliable, Efficient. In *AAAI*, 26631–26641. AAAI Press.

Meggendorfer, T.; Weininger, M.; and Wienhöft, P. 2025b. What Are the Odds? Improving Statistical Model Checking of Markov Decision Processes. In *QEST+FORMATS*, volume 16143 of *Lecture Notes in Computer Science*. Springer.

Morimoto, J.; and Atkeson, C. G. 2002. Minimax Differential Dynamic Programming: An Application to Robust Biped Walking. In *NIPS*, 1539–1546. MIT Press.

Munos, R. 2014. From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning. *Found. Trends Mach. Learn.*, 7(1): 1–129.

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

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust Adversarial Reinforcement Learning. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, 2817–2826. PMLR.

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

Schwartz, A. 1993. A Reinforcement Learning Method for Maximizing Undiscounted Rewards. In *ICML*, 298–305. Morgan Kaufmann.

Strehl, A. L. 2007. Model-Based Reinforcement Learning in Factored-State MDPs. In *Proceedings of the IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, 103–110.

Strehl, A. L.; and Littman, M. L. 2005. A theoretical analysis of Model-Based Interval Estimation. In *ICML*, volume 119 of *ACM International Conference Proceeding Series*, 856–863. ACM.

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.

Suilen, M.; Simão, T. D.; Parker, D.; and Jansen, N. 2022. Robust Anytime Learning of Markov Decision Processes. In *NeurIPS*.

Weissman, T.; Ordentlich, E.; Seroussi, G.; Verdú, S.; and Weinberger, M. J. 2003. Inequalities for the L_1 Deviation of the Empirical Distribution. Technical Report HPL-2003-97(R.1), Hewlett-Packard Laboratories.

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

Wolff, E. M.; Topcu, U.; and Murray, R. M. 2012. Robust control of uncertain Markov Decision Processes with temporal logic specifications. In *CDC*. IEEE.