





Certifiably Robust Policies for Uncertain Parametric Environments

Yannik Schnitzer(✉), Alessandro Abate, and David Parker

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

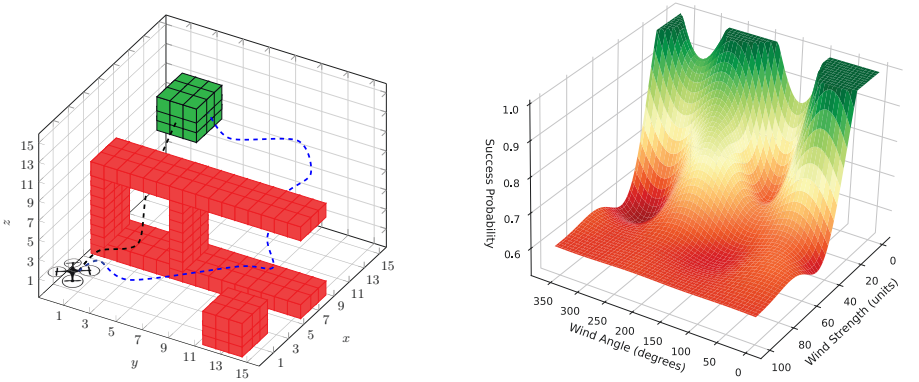
Abstract. We present a data-driven approach for producing policies that are provably robust across unknown stochastic environments. Existing approaches can learn models of a single environment as an interval Markov decision processes (IMDP) and produce a robust policy with a probably approximately correct (PAC) guarantee on its performance. However these are unable to reason about the impact of environmental parameters underlying the uncertainty. We propose a framework based on parametric Markov decision processes with unknown distributions over parameters. We learn and analyse IMDPs for a set of unknown sample environments induced by parameters. The key challenge is then to produce meaningful performance guarantees that combine the two layers of uncertainty: (1) multiple environments induced by parameters with an unknown distribution; (2) unknown induced environments which are approximated by IMDPs. We present a novel approach based on scenario optimisation that yields a single PAC guarantee quantifying the risk level for which a specified performance level can be assured in unseen environments, plus a means to trade-off risk and performance. We implement and evaluate our framework using multiple robust policy generation methods on a range of benchmarks. We show that our approach produces tight bounds on a policy's performance with high confidence.

1 Introduction

Ensuring the safety and robustness of autonomous systems in safety-critical tasks, such as unmanned aerial vehicles (UAVs), robotics or autonomous control, is paramount. A standard model for sequential decision making in these settings is a Markov decision process (MDP), which provides a stochastic model of the environment. However, real-world dynamics are complex, not fully known, and may evolve over time. Reasoning about *epistemic uncertainty*, which quantifies the lack of knowledge about the environment, can help construct *robust* policies that perform well across multiple possible stochastic environments.

Consider the UAV motion planning problem shown in Figure 1a, based on [6]. The goal is to navigate the drone safely to the target zone (green box) whilst avoiding obstacles (red regions). The drone's dynamics are influenced by weather conditions, such as wind strength or direction, potentially perturbing the drone from its intended route. These conditions can vary over time and may be difficult

© The Author(s) 2025
A. Gurfinkel and M. Heule (Eds.): TACAS 2025, LNCS 15698, pp. 63–83, 2025.
https://doi.org/10.1007/978-3-031-90660-2_4



(a) UAV motion planning environment with sample trajectories [6].

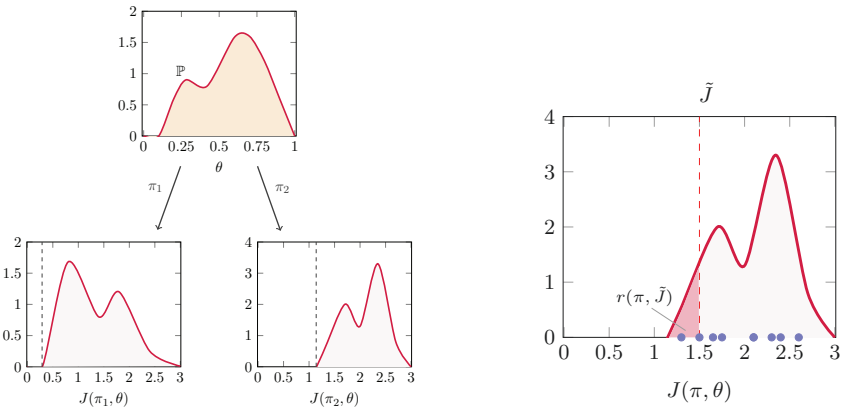
(b) Probability of task completion $J(\pi, \theta)$ for different parameter valuations θ.

Fig. 1: Example parametric environment with induced performance function.

to observe exactly. In low disturbance conditions (e.g., light wind), the drone can safely take the shorter route to the target (black dashed line). However, the drone should fly safely under all conditions, even if it flies overly cautiously in some. Therefore, a robust policy might take a detour through a less cluttered region of the environment (longer blue dashed line), ensuring a high probability of task completion even under more severe disturbances.

Epistemic uncertainty about the environment can be captured using *uncertain* MDPs, such as *interval MDPs* (IMDPs), which define a range of possible values for the probability of each transition between states of the model [21,53]. Under assumptions of independence, techniques such as robust dynamic programming [28,34] can then be used to efficiently generate *robust* policies for these IMDPs, i.e., policies that are optimal under *worst-case* assumptions about the true values of the transition probabilities. Furthermore, data-driven approaches, for example based on sampled trajectories through the environment, can be used to simultaneously learn both an IMDP and a robust policy for it [3,33,45,48], along with a probably approximately correct (PAC) guarantee on its performance.

In this paper, we present a general framework for synthesising provably robust policies in settings where environmental uncertainty is influenced by one or more *parameters*, e.g., wind strength/direction in the UAV example above. We model environments as *uncertain parametric MDPs* (upMDPs) [6], comprising a *parameter space* Θ of which each *parameter valuation* $\theta \in \Theta$ induces a standard MDP. Furthermore, an (unknown) distribution $\mathbb{P}$ over Θ represents the likelihood of each parameter valuation. Based on trajectories through multiple sampled instances of the environment, our goal is to produce *certifiably robust* policies. Instead of making worst-case assumptions across all possible parameter values, which can be overly conservative, we adopt a risk-based approach, providing a guaranteed level of performance for the policy with an associated *risk level* that



(a) Measure $\mathbb{P}$ and policy-induced densities over performance values $J(\pi_i, \theta)$.

(b) Performance guarantee $\tilde{J}$ and risk, estimated by sample performances (blue).

Fig. 2: For a fixed policy π, $J(\pi, \theta)$ is a random variable over performance values with measure $\mathbb{P}$ over valuations $\theta \in \Theta$ (left). We sample performances to bound the risk $r(\pi, \tilde{J})$, i.e., the probability for J to take a value less than $\tilde{J}$ (right).

quantifies the possibility of this level being violated. We also provide a tuning mechanism to adapt the trade-off between performance and risk.

To quantify the performance of a policy π in an MDP induced by parameter valuation $\theta \in \Theta$, we use an *evaluation function* $J(\pi, \theta)$. Typical examples include the probability to satisfy a specification expressed in temporal logics such as LTL [35] or PCTL [24] or an expected reward (see Section 2). For a fixed policy, J becomes a function in the valuations θ, as depicted in Figure 1b for the UAV example. When additionally considering the distribution $\mathbb{P}$ over parameter valuations, J becomes a random variable with respect to $\mathbb{P}$ describing the performance likelihood under policy π (see Figure 2a). Whereas the dashed vertical lines in Figure 2a indicate the worst-case performance of each policy, Figure 2b illustrates the risk measure $r(\pi, \tilde{J})$ that we use in this paper: the probability with which performance falls below a specified threshold $\tilde{J}$.

Deriving policies that are robust, i.e., which achieve high performance across either many or all possible environments, is a challenging problem. When Θ is finite, and assuming worst-case performance (i.e., ignoring $\mathbb{P}$), the model is referred to as a *multi-environment MDP*, for which finding an optimal robust memoryless policy is NP-hard, even for just two fully known environments [38]. For general upMDPs, recent work [39] finds robust policies but assumes that $\mathbb{P}$ can be sampled directly and that the resulting MDPs are fully known. In our setting, transition probabilities are unknown and are inferred from trajectories. This is comparable to the aforementioned work on PAC-learning of IMDPs [3,33,45,48], but these methods assume a single, fixed (but unknown) MDP.

In our work, there are *two layers* of uncertainty, resulting from (1) unknown parameter valuations inducing unknown MDPs, sampled from (2) the unknown

parameter distribution. In this setting, various learning-based methods known as *robust meta reinforcement learning* have been proposed [15,22,50]. Crucially, though, none of the existing learning algorithms are able to provide theoretical guarantees as to the performance of the generated policies in unseen environments; this is a core contribution of our framework.

An overview of our approach is illustrated in Figure 3. We assume access to multiple sampled environments $\mathcal{M}[\theta_i]$, each of which is an MDP induced by a parameter valuation θ_i from the unknown distribution $\mathbb{P}$. These are not fully known; instead we are able to access a set of sample trajectories from each one. In our UAV example, this equates to taking the drone outside on a new day and encountering a new set of environmental conditions (or a simulation of this).

Our framework divides the sample environments into two groups, a *training set* and a *verification* set. The training set is used to learn a robust policy. For this we build on existing IMDP-based policy learning methods and also consider robust meta reinforcement learning techniques. The verification set is used to derive the guarantees on the performance of the robust policy obtained from the training set. For this, we apply PAC IMDP learning to each unknown MDP in the verification set. Concretely, we use sample trajectories to infer IMDP overapproximations, which contain the true, unknown MDPs with a user-specified confidence. From those, we can derive lower bounds on the performance J of the learned policy in each of the environments, which hold with the specified confidence.

To tackle the higher layer of uncertainty and infer a bound for the policy's robust performance over the entire unknown distribution $\mathbb{P}$ underlying the sample MDPs, we develop a new approach based on *scenario optimisation* [11,12]. This takes samples of the performance J and a user-specified performance bound $\tilde{J}$ and provides a PAC guarantee on the probability of the performance on a new sample being less than $\tilde{J}$, i.e., the *risk*. However, in our setup we do not obtain samples of J directly, but derive lower bounds from the learned PAC IMDPs, which only hold up to a certain confidence. Our key theoretical contribution, presented in Section 3.3, is a generalisation of the scenario approach that can handle samples whose values are only known to lie in a confidence interval.

Our theoretical results combine the two layers of uncertainty: (1) the finite sampling of MDPs from the distribution $\mathbb{P}$, (2) the fact that sampled MDPs are unknown, so the performances of the learned policy are only inferable up to a certain confidence. The result is a single PAC guarantee on the policy's performance which holds with a high, user-specified confidence.

Furthermore, our framework allows tuning of the trade-off between performance guarantee and risk. By excluding the k worst-case sample environments, users can discard unlikely outliers, resulting in a higher performance guarantee at the cost of an increased risk bound, adjustable to the level the user considers admissible. We implement our framework as an extension of the PRISM model checker [31] and show that it can tightly quantify the performance and associated risk of learned policies on a range of benchmarks.

In summary, our contributions are: (1) a novel framework and techniques for producing certifiably robust policies in uncertain parametric MDPs for which

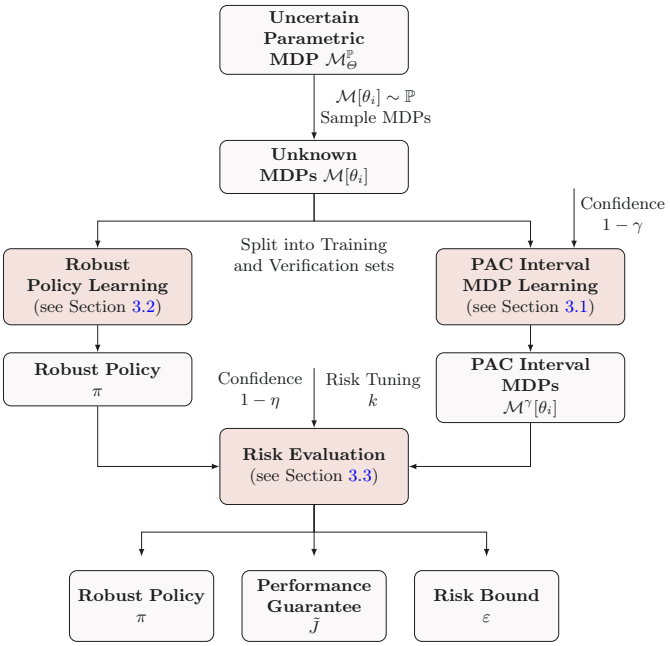


Fig. 3: Overview of our framework to derive performance and risk guarantees for policies learned on upMDPs. The setup includes two layers of uncertainty: we sample and analyse unknown environments from an unknown distribution.

both the parameters and transition probability functions are unknown; (2) new theoretical results which yield PAC guarantees on a policy's robust performance on unseen environments, where sample environments are unknown and can only be estimated from trajectories; (3) an implementation and evaluation of the framework on a range of benchmarks.

An extended version of this paper, with full proofs of all results, extended experiments and further technical details, can be found in [42].

1.1 Related Work

Epistemic uncertainty in MDPs has received broad attention across many areas, including formal methods, planning and reinforcement learning [5]. As mentioned above, there are various ways to model this using uncertain MDPs [21,53]. There are also techniques such as robust dynamic programming to synthesise robust policies for these models [28,34] and approaches to learn the models from trajectory data [3,33,45,48]. In this work, however, we investigate *parametric* uncertainty sets with unknown distributions over parameter valuations.

Uncertain parametric MDPs have emerged as a common model in meta reinforcement learning [15,19,20,22,23] and gained attention in formal methods [6,39]. On the one hand, meta reinforcement learning trains policies on multiple unknown environments sampled from an upMDP, using policy gradient methods [49], in

order to generalise to unseen environments. However, to our knowledge, none of these algorithms provide theoretical generalisation guarantees, either on their average [19,23] or their robust performance [15,22].

On the other hand, existing formal methods approaches to upMDPs do not offer the generality of meta RL setups. The work in [6] uses scenario methods and provides PAC guarantees, but for the existence of a policy that achieves a certain performance, not robust policy synthesis; they also require full knowledge of sampled parameter valuations and environments. In [39], concrete robust policies are synthesised with a PAC guarantee for performance on unseen environments, but this also relies on complete knowledge of all sampled valuations, reducing it to a special case of our approach. In our work, we address the very general problem of unknown sample environments; we target the scalability and generality of meta RL, while providing formal guarantees that are independent of the model size and previously unattainable for policy training methods like those in [39].

Also related is [16] which uses parametric MDPs in a Bayesian setting; parameter valuations are unknown but the model's transition functions are known and assumed to be defined by polynomial expressions. Other recent work in [14] combines, like us, the scenario approach and parametric MDPs, but for a different setting that assumes uniform distributions over parameter spaces. We also mention [7], which uses scenario optimisation with imprecise constraints to analyse continuous-time Markov chains with uncertain transition rates.

2 Preliminaries

We review the key formalisms used in our approach. Let $\Delta(S) = \{\mu \colon S \to [0,1] \mid \sum_s \mu(s) = 1\}$ denote the set of all probability distributions over a finite set S.

Definition 1 (Parametric MDP). *A* parametric Markov decision process *(pMDP) is a tuple $M_\Theta = (S, s_I, A, P_\Theta)$, where S and A are finite state and action spaces, $s_I \in \Delta(S)$ is the initial state distribution, and $P_\Theta \colon \Theta \times S \times A \to \Delta(S)$ is the parametric transition probability function over the parameter space Θ. Fixing a valuation $\theta \in \Theta$ induces a standard MDP $\mathcal{M}_\Theta[\theta]$, or $\mathcal{M}[\theta]$ for short, with transition kernel $P_\theta \colon S \times A \to \Delta(S)$ defined as $P_\theta(s,a,s') = P_\Theta(\theta,s,a,s')$.*

Parametric MDPs can be seen as an abstract model for a *set* of MDPs, i.e., they represent the instantiations induced by all possible valuations $\theta \in \Theta$. They are closely related to the model class of uncertain MDPs [34,53], in which each transition is associated with (potentially interdependent) sets of possible values.

Definition 2 (Uncertain Parametric MDP). *An* uncertain parametric Markov decision process *(upMDP) $\mathcal{M}^{\mathbb{P}}_\Theta = (\mathcal{M}_\Theta, \mathbb{P})$ is a pMDP $\mathcal{M}_\Theta$ with a (potentially unknown) probability measure $\mathbb{P}$ over the parameter space Θ.*

We assume that upMDPs are *graph preserving*, meaning that induced MDPs share a common topology: $\forall s, s' \in S, a \in A \colon (\forall \theta \in \mathrm{supp}(\mathbb{P}) \colon P_\theta(s,a,s') = 0) \vee (\forall \theta \in \mathrm{supp}(\mathbb{P}) \colon P_\theta(s,a,s') > 0)$. Although not strictly required, this assumption

can be crucial for efficiently solving learned IMDP approximations of the induced MDPs [33,53]. We describe in Section 3.4 how to lift this assumption by resorting to techniques which only approximate the performance and not the model.

Policies resolve action choices in MDPs and upMDPs. They are mappings $\pi\colon (S \times A)^* \times S \to \Delta(A)$, assigning a distribution over actions based on (finite) histories of states and actions. In this work, we focus on synthesising memoryless policies $\pi\colon S \to \Delta(A)$. While our theoretical results apply to arbitrary policy classes, learning and evaluating more expressive policies, such as those with finite memory, can be computationally expensive. We elaborate on this in Section 3.4.

Definition 3 (Evaluation Function). *For an upMDP $\mathcal{M}_\Theta^{\mathbb{P}} = (\mathcal{M}_\Theta, \mathbb{P})$, an* evaluation function $J\colon \Pi \times \Theta \to \mathbb{R}$ *maps a policy π and a parameter valuation θ to a performance value. We also denote by $J(\pi, \mathcal{M})$ the evaluation of policy π on an arbitrary MDP $\mathcal{M}$, such that $J(\pi, \theta) = J(\pi, \mathcal{M}[\theta])$.*

Typical evaluation functions include the reachability probability $\mathbf{Pr}^\pi_{\mathcal{M}}(\Diamond T)$ of eventually reaching a set of target states $T \subseteq S$, the reach-avoid probability $\mathbf{Pr}^\pi_{\mathcal{M}}(\neg C \,\mathcal{U}\, T)$ of reaching states in T while not entering a set of avoid states $C \subseteq S$, the expected reward $\mathbb{E}^\pi_{\mathcal{M}}(\Diamond T)$ accumulated before reaching a set T, given a reward structure, expected accumulated reward over finite or infinite time horizons, or more sophisticated properties expressed in temporal logics such as LTL [35] or PCTL [24]. Throughout this work's technical presentation, we assume that performance J is maximised. By changing the directions of optimisation and inequalities, our results also directly apply to the dual minimisation case.

For a fixed policy π and upMDP $\mathcal{M}_\Theta^{\mathbb{P}}$, the evaluation function J only depends on the valuations θ. Hence, J is a random variable with measure $\mathbb{P}$ (see Figure 2a). The *violation risk* is the probability that policy π achieves a value less than a stated performance guarantee $\tilde{J}$ on $\mathcal{M}_\Theta^{\mathbb{P}}$ (see Figure 2b).

Definition 4 (Violation Risk). *The* violation risk *of policy π for performance guarantee $\tilde{J} \in \mathbb{R}$, denoted by $r(\pi, \tilde{J})$, is defined as:*

$$r(\pi, \tilde{J}) = \mathbb{P}\left\{\theta \in \Theta\colon J(\pi, \theta) < \tilde{J}\right\}. \tag{1}$$

There is an inherent trade-off between violation risk and performance guarantee: a higher guarantee is associated with a higher risk, regardless of $\mathcal{M}_\Theta^{\mathbb{P}}$ and J.

The framework we present in this paper is based on learning and solving approximations of MDPs in the form of *interval MDPs* [21,53].

Definition 5 (Interval MDP). *An interval Markov decision process (IMDP) $\mathcal{M}^I = (S, s_I, A, P^I)$ is an MDP with a probabilistic interval transition function $P^I\colon S \times A \times S \to \mathbb{I}$, where $\mathbb{I} = \{[a, b] \mid 0 < a \leq b \leq 1\} \cup \{[0, 0]\}$ is the set of all graph-preserving intervals. We say that IMDP $\mathcal{M}^I$* includes *MDP $\mathcal{M}$, denoted by $\mathcal{M} \in \mathcal{M}^I$, if $\mathcal{M}$ and $\mathcal{M}^I$ share the same state and action spaces, and $P(s, a, s') \in P^I(s, a, s')$ for all $s, s' \in S, a \in A$.*

For IMDPs, we typically adopt a *robust* view of a policy's performance, i.e., the *worst-case* (minimum) value over any included MDP. We lift the evaluation function J to an IMDP $\mathcal{M}^I$ as follows:

$$J(\pi, \mathcal{M}^I) = \min_{\mathcal{M} \in \mathcal{M}^I} J(\pi, \mathcal{M}) \implies J(\pi, \mathcal{M}^I) \leq J(\pi, \mathcal{M}) \text{ for all } \mathcal{M} \in \mathcal{M}^I. \quad (2)$$

For key classes of properties used here (e.g., reachability probabilities, rewards), this value can be obtained via robust dynamic programming [28,34,53].

3 Learning Certifiably Robust Policies for upMDPs

This section presents our framework for computing performance and risk guarantees for learned policies in uncertain parametric MDPs. An overview of this framework was illustrated in Figure 3 and introduced in Section 1. We assume a fixed upMDP $\mathcal{M}^{\mathbb{P}}_{\Theta}$ and access to a sample set $\mathcal{D} = \{\mathcal{M}[\theta_i] \mid \theta_i \sim \mathbb{P}\}$ of unknown MDPs. This sample set is split into disjoint training and verification sets. The training set is used to compute a robust policy π, which is then evaluated on the verification set to derive a performance guarantee $\tilde{J}$ and bound the violation risk when the policy is deployed in an unseen environment sampled from $\mathbb{P}$. The overall goal is formally stated as follows.

Problem 1 *Given a upMDP $\mathcal{M}^{\mathbb{P}}_{\Theta}$ with unknown parameter distribution $\mathbb{P}$, an evaluation function J, and a confidence level $\eta > 0$, find a robust policy π, a (tight) performance guarantee $\tilde{J}$, and a (tight) risk bound $\varepsilon > 0$, such that:*

$$\Pr\left\{r(\pi, \tilde{J}) \leq \varepsilon\right\} \geq 1 - \eta.$$

There exist trivial solutions to Problem 1, such as selecting a very high risk bound or a very low performance guarantee. We aim to identify a tight solution that maximizes the performance guarantee with a precise, low risk bound.

The core part of our framework is the means to establish these performance and risk bounds for policies evaluated in unknown environments. From each unknown MDP in the verification set, we sample trajectories to learn an IMDP approximation that includes the MDP with high confidence. Solving these IMDPs yields probabilistic bounds on the policy's performance in each environment.

Our main theoretical results build upon scenario optimisation [11,12], the principal challenge being to incorporate the additional layer of uncertainty introduced by only being able to estimate the policy's performance in each unknown sampled environment. The result is a single PAC guarantee on the policy's performance in unseen environments, stated in Problem 1 above.

The process of establishing these guarantees is agnostic to the manner in which policies are produced. We consider two approaches, first taking a novel combination of existing methods for IMDPs and upMDPs [39,48], and secondly adopting a gradient-based technique from robust meta reinforcement learning [22].

The remainder of the section is structured as follows. Since PAC learning of IMDPs is used in multiple places, we discuss this first, in Section 3.1. Section 3.2 covers robust policy learning, Section 3.3 presents our main theoretical contributions and Section 3.4 describes several optimisations and extensions.

3.1 PAC IMDP Learning of Unknown MDPs

We follow established approaches for PAC learning of IMDP approximations introduced in [3,33,45,48]. Consider an unknown MDP $\mathcal{M}[\theta_i]$. We assume access to trajectories from $\mathcal{M}[\theta_i]$, consisting of sequences of triples (s, a, s') representing states, chosen actions and successor states. Leveraging the Markov property of MDPs, we treat each triple as an independent Bernoulli experiment to estimate the transition probability to state s' from s when choosing action a. We denote the number of times action a was chosen in state s across all sample trajectories as $\#(s, a)$, and the number of times this choice led to s' as $\#(s, a, s')$. The resulting point estimate for the transition probability is thus given by:

$$\tilde{P}(s, a, s') = \frac{\#(s, a, s')}{\#(s, a)}, \tag{3}$$

for $\#(s, a) > 0$. We construct an IMDP by transforming the point estimates into PAC confidence intervals [10]. Traditionally, this is achieved with Hoeffding's inequality [27,45]. Recent work has demonstrated that significantly tighter model approximations can be obtained by employing inequalities tailored to the binomial distribution, such as the Wilson score interval with continuity correction [33,54].

Let $1 - \gamma$ with $\gamma > 0$ be the desired confidence level for $\mathcal{M}[\theta_i]$ to be included in the IMDP, which we denote $\mathcal{M}^\gamma[\theta_i]$. This confidence is distributed over all n_u unknown transitions as $\gamma_P = \gamma / n_u$. Let $H = \#(s, a)$ and $\tilde{p} = \tilde{P}(s, a, s')$. For each unknown transition, the transition probability interval is given by:

$$P^\gamma(s, a, s') = [\max(\mu, \underline{p}_{wcc}), \min(\overline{p}_{wcc}, 1)], \tag{4}$$

with:

$$\underline{p}_{wcc} = \left(2H\tilde{p} + z^2 - z\sqrt{z^2 - \frac{1}{H} + 4H\tilde{p}(1-\tilde{p}) + 4\tilde{p} - 2} - 1\right) / \left(2(H + z^2)\right), \tag{5}$$

$$\overline{p}_{wcc} = \left(2H\tilde{p} + z^2 + z\sqrt{z^2 - \frac{1}{H} + 4H\tilde{p}(1-\tilde{p}) - 4\tilde{p} + 2} + 1\right) / \left(2(H + z^2)\right), \tag{6}$$

where z is the $1 - \frac{\gamma_P}{2}$ quantile of the standard normal distribution [33] and $\mu > 0$ is an arbitrarily small quantity to preserve the known graph structure. For unvisited state action pairs with $\#(s, a) = 0$, we set $P^\gamma(s, a, s') = [\mu, 1]$, for all s' in the known support. $P^\gamma(s, a, s')$ contains the true transition probability $P(s, a, s')$ with a confidence of at least $1 - \gamma_P$. By applying a union bound over the unknown transitions, we obtain the following overall guarantee:

Lemma 1 ([33]). *The true, unknown MDP $\mathcal{M}[\theta_i]$ is contained in its IMDP overapproximation $\mathcal{M}^\gamma[\theta_i]$ with probability at least $1 - \gamma$.* □

The confidence in the approximation of each environment is independent of the number or length of the trajectories analysed. However, more or longer trajectories generally lead to higher state-action counts, resulting in tighter intervals. In Section 4, we examine how the number of trajectories analysed influences the tightness of our performance guarantee and the associated risk.

3.2 Robust Policy Learning

We consider two distinct approaches to robust policy learning: *robust IMDP policy learning* and *robust meta reinforcement learning*. For the former, we propose a combination of techniques for robust policy synthesis for upMDPs with access to fully known sample environments [39] and IMDP learning for single unknown environments [48]. For the latter, we adopt a class of algorithms that optimise a policy's robust performance using policy gradient-methods [15,22].

Robust IMDP Policy Learning. Similarly to PAC IMDP learning in Section 3.1, we use sample trajectories to compute an IMDP overapproximation [48] for each unknown MDP in the training set. Then, like in [39], we combine models across the training set to perform policy synthesis. To obtain a policy that is robust across all samples, the learned IMDPs are combined by merging the transition intervals of each IMDP as $[a, b] \sqcup [c, d] = [\min(a, c), \max(b, d)]$. The resulting IMDP over-approximates all training MDPs, and the corresponding optimal deterministic policy considers the worst-case probability for each transition [53]. As the IMDPs for the training set are only used for policy synthesis and not for formal risk or performance analysis, we are not restricted to PAC IMDP learning. We can leverage a rich pool of interval learning algorithms, which, while lacking formal inclusion guarantees, provide empirically tighter intervals from fewer trajectories. A detailed overview and comparison of interval learning methods and their model approximation capabilities is conducted in [48]. We evaluate the best-performing approaches in our benchmarks in Section 4.

Robust Meta Reinforcement Learning. IMDP policies can be overly conservative, as they consider the worst-case scenario for each transition independently [39]. Additionally, IMDP learning produces memoryless deterministic policies. While sufficient for optimality in IMDPs, there exist upMDPs where an optimal robust policy requires randomization or memory [38].

Robust meta-reinforcement learning (RoML) extends classical reinforcement learning from a single MDP to upMDPs [15,22]. RoML employs policy gradient methods [49] to optimize a policy's performance across sampled training environments, estimating performance from sampled trajectories in each unknown environment. Unlike standard meta-RL, which maximises expected rewards across environments [9] and often yields strong average but poor worst-case performance, RoML prioritises robustness. It trains a policy using RL techniques to optimise the performance in the worst-case environment (via the max-min objective) [15] or in the α-quantile of worst-case environments (via a risk-aware CVaR objective) [22]. RoML differs from *robust RL* [18,52], which focuses on a single uncertain MDP, seeking a policy that is robust to its internal uncertainty. In contrast, RoML

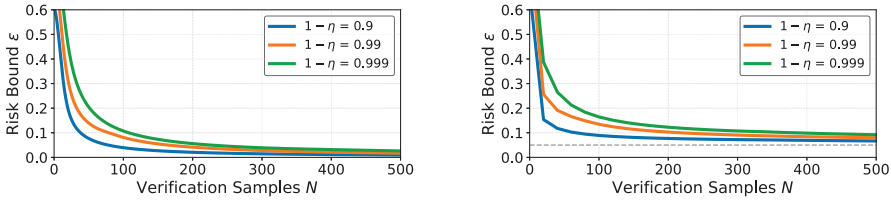


Fig. 4: Example risk bounds obtained from Theorem 1 (left) and Theorem 2 (right) for IMDP confidence $\gamma = 10^{-4}$. For Theorem 2, 5% of samples are discarded.

addresses multiple MDPs sampled from an unknown distribution, aiming for a policy that generalizes to the full distribution by achieving a strong robust performance across the sampled MDPs. This aligns well with our setup, and we refer the reader to [15,22] for further details on this approach.

3.3 Certifying Policy Performance and Risk in upMDPs

We now present our main theoretical contributions for quantifying the performance and violation risk of a policy π deployed in unseen environments of an upMDP. Specifically, we provide two results that derive PAC guarantees from lower bounds on π's performance, which we obtain by building and analysing PAC IMDPs (see Section 3.1) for the sampled environments that make up the verification set.

The first result extends the scenario approach [11,12] to incorporate additional uncertainty, as the lower bounds hold only with a certain confidence. This provides a PAC guarantee on the policy's performance, reasoning over the unknown distribution of true environments $\mathbb{P}$, based solely on estimations from sampled unknown environments attained by IMDP learning. The second result introduces flexibility in balancing the risk-performance trade-off. By extending a second result of the scenario approach, we allow the performance bound to be tuned by excluding worst-case outlier samples from the analysis, potentially leading to a higher performance guarantee at the cost of an increased risk bound.

Assume that the verification set comprises N sampled MDPs, from which we have learnt the PAC IMDPs $\{\mathcal{M}^\gamma[\theta_i]\}_{1\leq i\leq N}$. This establishes probabilistic lower bounds on the policy π's performance in the underlying unknown MDPs. From Equation (2), we have that $\mathcal{M}[\theta_i] \in \mathcal{M}^\gamma[\theta_i] \Rightarrow J(\pi, \mathcal{M}^\gamma[\theta_i]) \leq J(\pi, \mathcal{M}[\theta_i])$, where $J(\pi, \mathcal{M}^\gamma[\theta_i])$ can be obtained by standard solution methods for IMDPs, such as robust dynamic programming [28,34]. By Lemma 1, it follows that:

$$\mathbb{P}\{J(\pi, \mathcal{M}^\gamma[\theta_i]) \leq J(\pi, \mathcal{M}[\theta_i])\} \geq 1 - \gamma. \tag{7}$$

To bound the violation risk, we formulate the problem as a convex optimisation problem with randomised constraints—a *scenario program* [13]. We extend the generalisation theory of the *scenario approach* [11] to account for the uncertainty in the lower performance bounds (see Equation (7)). The detailed formulation and derivation of our generalisation can be found in [42].

Theorem 1. *Given N i.i.d. sample MDPs $\mathcal{M}[\theta_i]$ and IMDPs $\mathcal{M}^\gamma[\theta_i]$, such that $\mathbb{P}\{\mathcal{M}[\theta_i] \in \mathcal{M}^\gamma[\theta_i]\} \geq 1-\gamma$. For any policy π and confidence level $1-\eta$, with $\eta > 0$, it holds that:*

$$\mathbb{P}^N \left\{ r(\pi, \tilde{J}^\gamma) \leq \varepsilon(N, \gamma, \eta) \right\} \geq 1 - \eta, \tag{8}$$

where $\tilde{J}^\gamma = \min_i J(\pi, \mathcal{M}^\gamma[\theta_i])$, and $\varepsilon(N, \gamma, \eta)$ is the solution to the following, for any $K \leq N$:

$$\sum_{i=K}^{N} \binom{N}{i} (1-\gamma)^i \gamma^{N-i} - (1-\eta) = \sum_{i=0}^{N-K} \binom{N}{i} \varepsilon^i (1-\varepsilon)^{N-i}. \tag{9}$$

Proof sketch. The standard scenario approach in one dimension considers a set of i.i.d. performance samples $J_1, \ldots, J_N$ and provides a bound on the probability that the next sampled performance is lower than the minimum. In our case, we only have lower bounds on the actual performances, with $\mathbb{P}\{J_i^\gamma \leq J_i\} \geq 1-\gamma$. Assuming all lower bounds are valid with probability $(1-\gamma)^N$, we obtain an under-approximation of the true solution to the scenario program, and the union bound combines the confidences of the scenario approach and the validity of the lower bounds. However, this confidence becomes very small for large sample sizes N, even when $1-\gamma$ is close to 1. Conversely, with small sample sizes, the overall confidence remains low due to the weaker scenario confidence. To mitigate this, we require only K of the N lower bounds to be valid, which holds with high probability for small values of γ, even when K is close to N. As a result, we can exclude a small number $N-K$ of samples, assuming them to be violated under the scenario approach, thereby soundly over-approximating the risk. This only marginally increases the stated risk bound while significantly reducing the confidence overhead. Since we cannot specify which bounds are valid, we use the minimum over all lower performance bounds as an under-approximation of the solution to all scenario sub-programs with $N-K$ discarded constraints, providing a sound performance guarantee. The complete proof, including detailed bounds and derived inequalities, can be found in [42].

Theorem 1 bounds the risk for a policy to achieve a performance less than the minimum performance on any of the IMDPs. This bound only depends on values we can observe from the learned IMDP approximations. The theorem includes a tuning parameter $K \leq N$. The bound is valid for any value of K, and to obtain the tightest bound, we enumerate possible values and solve the resulting equation. For a fixed K, the left-hand side of Equation (9) is constant, and the right-hand side is the cumulative distribution function of a beta distribution with $K+1$ and $N-K$ degrees of freedom [13], which is easy to solve numerically for its unique solution in the interval $[0, 1]$ using bisection [6,40]. To the best of our knowledge, this is the first result to establish PAC guarantees on policy performance in unseen environments of upMDPs, in a setting where sample environments are unknown and can only be estimated from trajectories. Figure 4 illustrates the resulting risk bounds for example values. We assess the quality and tightness of our performance and risk bounds in the benchmarks presented in Section 4.

We extend Theorem 1 to allow tuning of the risk-performance trade-off by *discarding* samples [12]. Instead of bounding the risk for the policy to achieve a performance less than the minimum, we state a bound for the kth order statistic of the verification set. Users can choose k for a permissible risk level and a potentially higher performance guarantee, avoiding constraints from samples in the unlikely tail of the distribution.

Definition 6 (Order Statistic). *For a set of N samples $J_1, \ldots, J_N \in \mathbb{R}$ and $0 \leq k < N$, the kth order statistic $\tilde{J}_{(k)}$ is the kth smallest element when arranging all samples from smallest to largest.*

Theorem 2. *Given N i.i.d. sample MDPs $\mathcal{M}[\theta_i]$ and IMDPs $\mathcal{M}^\gamma[\theta_i]$, such that $\mathbb{P}\{\mathcal{M}[\theta_i] \in \mathcal{M}^\gamma[\theta_i]\} \geq 1-\gamma$, for any policy π, confidence level $1-\eta$ with $\eta > 0$, and number k of discarded samples, it holds that*

$$\mathbb{P}^N \left\{ r(\pi, \tilde{J}^\gamma_{(k)}) \leq \varepsilon_{(k)}(N, \gamma, \eta) \right\} \geq 1 - \eta, \tag{10}$$

where $\tilde{J}^\gamma_{(k)}$ is the kth order statistic of the performances $J(\pi, \mathcal{M}^\gamma[\theta_i])$, and $\varepsilon_{(k)}(N, \gamma, \eta)$ is the solution to the following, for any $K \leq N - k$:

$$\sum_{i=K}^{N-k} \binom{N-k}{i} (1-\gamma)^i \gamma^{N-k-i} - (1-\eta) = \sum_{i=0}^{N-K} \binom{N}{i} \varepsilon^i (1-\varepsilon)^{N-i}. \tag{11}$$

□

When $k = 0$ and no samples are discarded, Theorem 2 specialises to Theorem 1. The proof is an extension incorporating the additional uncertainty of the lower bounds obtained from the PAC IMDPs into the *sampling-and-discarding* theorem from scenario optimisation [12]. We detail the derivation of the bound in [42] and analyse its tightness in the experiments in Section 4.

3.4 Optimisations and Extensions

Finally in this section, we present some optimisations and extensions for our approach. First we show that, if there is additional knowledge as to the parametric structure of the upMDP, we can leverage this to obtain tighter approximations of the sample environments. Conversely, we describe how to seamlessly apply our framework and results to setups with *less* model knowledge, i.e., where not even the graph structure nor the (possibly infinite) state space is known. Furthermore, we outline how our results apply to more general setups where parameters influence not only transition probabilities, but also the evaluation functions, i.e., the specifications or *tasks* may vary across samples, aligning it with the setup commonly considered in meta reinforcement learning [15,22].

Model-based Optimisations. IMDP learning as described in Section 3.1 requires no knowledge of an MDP beyond its graph structure. However, additional information about the environment can yield tighter approximations with fewer samples. In cases where certain parameters, like temperature or air pressure, and their effect on some transition probabilities are known exactly, those transitions

can be treated as singleton intervals. This reduces the need for approximation and decreases the number of learned transitions n_u.

Additionally, we can apply *parameter tying* [36,37] to parameters appearing across different transitions. For instance, consider two transitions sharing the same parameterisation, $P_\Theta(s,a,s') = P_\Theta(t,b,t')$. We can combine the counts from both transitions since they represent the same Bernoulli experiment. Let $sim(s,a,s') = \{(t,b,t') \mid P_\Theta(s,a,s') = P_\Theta(t,b,t')\}$ denote the set of transitions with identical parametrisation. By plugging the combined counts, $\#^T(s,a,s') = \sum_{(t,b,t')\in sim(s,a,s')} \#(t,b,t')$ and $H^T(s,a,s') = \sum_{(t,b,t')\in sim(s,a,s')} \#(t,b)$ into Equations (3) and (4), we can obtain a tighter interval for both transitions.

Our experiments in Section 4 use model-based optimisations and the full evaluation in [42] compares the results with and without optimisations.

Statistical Model Checking. To approximate the performance of a learned policy in unknown sample environments, our framework is not limited to PAC IMDP learning. Various forms of statistical model checking (SMC) [1,26,32] can be applied, as long as they provide a lower bound J_i^γ on a policy's performance in a single environment induced by parameters θ_i, with a formally quantified confidence $\Pr\{J(\pi,\theta_i) \geq J_i^\gamma\} \geq 1-\gamma$. SMC techniques that require less information than PAC IMDP learning include those that rely on the minimum probability p_{min} potentially present in the model to infer the MDP's end-components with the desired confidence [3,17,33], or those that operate in a fully black-box setting with no model knowledge, approximating only the performance value. However, the latter techniques are typically restricted to finite-horizon properties [43,56].

Uncertain Specifications or Tasks. The meta reinforcement learning literature usually considers upMDPs where both transition probabilities and reward structure depend on parameters [19,22,23]. Our framework encompasses this problem class, and our results carry across directly. Parameterised rewards or specifications can be integrated into the evaluation function J, allowing parameters to affect both transitions and rewards. While the formal methods community has explored uncertain rewards and specifications to a lesser extent [4,41,44], we believe this is a promising direction for future work, particularly in extending PAC guarantees to uncertain specifications and objectives.

Non-memoryless Policies. In this work, we focus on synthesis of *memoryless* policies, which are sufficient for optimal performance under many common performance functions, such as reachability and reach-avoid specifications, in both single MDPs and IMDPs. However, for multiple environments or more complex specifications, such as general linear-time (LTL) properties, an optimal robust policy may require memory [34,38,55]. Our theoretical results in Theorems 1 and 2 apply to arbitrary policy classes, provided that policy performance can be evaluated on the learned IMDP approximations to obtain sample performance values. But solution of IMDPs in these cases presents challenges. For example, the standard automaton product constructions can be used to find finite-memory policies that optimise LTL specifications [55] but, in addition to increased compu-

Table 1: Salient characteristics of the evaluated benchmarks.

Benchmark	Evaluation J	Opt.	#Parameters	#States	#Transitions
UAV [6]	$\mathbf{Pr}(\neg C\,\mathcal{U}\,T)$	max	15	4096	86912
Aircraft Collision [30]	$\mathbf{Pr}(\neg C\,\mathcal{U}\,T)$	max	2	303	3468
Firewire [25]	$\mathbf{Pr}(\Diamond T)$	min	1	80980	112990
Semi-Auton. Vehicle [29]	$\mathbf{Pr}(\Diamond T)$	max	2	411	1503
Betting Game [8]	$\mathbb{E}(\Diamond T)$	max	1	480	2730
Chain [2]	$\mathbb{E}(\Diamond T)$	min	1	7	42

tational cost, the addition of memory means moving from a static to a dynamic uncertainty model [47,55], yielding overly conservative performance bounds.

4 Experimental Evaluation

We implemented our framework as an extension of the PRISM model checker [31], which provides trajectory generation and (robust) value iteration for MDPs and IMDPs[1]. We have evaluated our approach on a range of established benchmark environments used in similar work: an Aircraft Collision Avoidance [30], a Chain Problem [2], a Betting Game [8], a Semi-Autonomous Vehicle [29,46], the Firewire protocol [25], and the previously mentioned UAV [6]. Table 1 shows the salient features of the environments. We provide detailed descriptions of each benchmark, including the parameters and their distributions $\mathbb{P}$ in [42].

Setup. Since our approach is the first to provide policy performance guarantees under two layers of uncertainty, i.e., unknown sample environments from an unknown distribution, our experiments focus on assessing the quality of these guarantees. We study: (1) the tightness of the performance level $\tilde{J}$, assessing how closely our stated robust performance guarantee derived from the learned PAC IMDPs aligns with the actual robust performance on the true underlying sample MDPs that are hidden from the algorithm; (2) the quality of the risk bound ε derived from Theorems 1 and 2 with respect to the true violation risk $r(\pi, \tilde{J})$.

Our approach includes two sampling dimensions corresponding to the two layers of uncertainty: (1) the number of unknown MDPs induced by parameter valuations sampled from the distribution $\mathbb{P}$; (2) the number of sample trajectories generated in each unknown MDP. For the first dimension we consider a total of 600 sample MDPs, which we divide equally into training and verification sets. For the Firewire benchmark, we consider 150 verification samples. The second dimension is evaluated for up to 10^6 trajectories in each sampled environment.

For policy learning, we consider the two methods described in Section 3.2: robust IMDP policy learning and robust meta reinforcement learning with the max-min objective, implemented using the *Gymnasium* Python framework [51]. This illustrates the applicability of our approach to distinct state-of-the-art policy learning algorithms. We focus here on guarantees, rather than comparing policy learning methods, but we include statistics for both in [42] and refer the reader to, e.g., [15,22,48] for an in-depth comparison of methods.

[1] The implementation is available at: https://doi.org/10.5281/zenodo.14717176.

Table 2: Resulting performances, guarantees and risk bounds.

Benchmark	Performance J	Guarantee $\tilde{J}$	Risk Bound ε	Empirical Risk $r(\pi, \tilde{J})$	Risk Bound $\varepsilon_{(5)}$/Empirical Risk	Risk Bound $\varepsilon_{(10)}$/Empirical Risk	Runtime per 10^4 trajectories
UAV	0.7110	0.7100	0.027	0.003	0.052 / 0.023	0.075 / 0.057	1.51s
Aircraft Collision	0.5949	0.5900	0.027	0.004	0.052 / 0.017	0.075 / 0.046	0.35s
Firewire	0.1946	0.1967	0.055	0.004	0.103 / 0.039	0.146 / 0.081	14.9s
Semi-Auton. Vehicle	0.7854	0.7767	0.027	0.004	0.052 / 0.018	0.075 / 0.033	0.50s
Betting Game	30.78	30.65	0.027	0.005	0.052 / 0.016	0.075 / 0.026	1.12s
Chain	127.2	128.0	0.027	0.002	0.052 / 0.032	0.075 / 0.054	0.32s

For producing guarantees, we set the inclusion confidence level for the PAC IMDPs learned on the verification set to $\gamma = 10^{-4}$ and the overall confidence when applying Theorems 1 and 2 to $\eta = 10^{-2}$. Optimisations from Section 3.4 are applied (see [42] for results of their impact). All experiments were conducted on a 3.2 GHz Apple M1 Pro CPU with 10 cores and 16 GB of memory.

Results. Table 2 summarises the resulting performances and guarantees for the best-performing policy. All results are obtained after processing the full set of trajectories. We first give the *true robust performance* J, i.e., the policy's performance on the worst-case true MDP, which is hidden from the algorithm. We then report the key outputs of our approach: the *robust performance guarantee* $\tilde{J}$, representing the worst-case performance on the learned PAC IMDPs, and the *risk bound* ε for the performance guarantee $\tilde{J}$, obtained using Theorem 1. We also show an empirical estimate of the *true violation risk* $r(\pi, \tilde{J})$, computed over 1000 fresh sample MDPs. To evaluate the bounds derived via Theorem 2, we include the risk bounds $\varepsilon_{(k)}$ for discarding $k = 5$ and $k = 10$ worst-case samples, alongside estimates of the corresponding true violation risks.

For example, in the UAV case study, the table shows that the learned policy achieves at least $J = 0.711$ probability of reaching the goal without crashing into an obstacle on any sampled true MDP hidden from the algorithm. The learned IMDP approximations certify a minimum performance of $\tilde{J} = 0.71$, with the probability of performing worse on a fresh sample MDP bounded by $\varepsilon = 0.027$. On 1000 fresh samples, the policy actually performed worse in only 0.3% of cases.

Figure 5 shows the learning process and the derived performance guarantee for the Betting Game benchmark. We plot the true robust performance J (solid line), and the robust performance guarantee $\tilde{J}$ (dashed line) against the number of trajectories processed for each unknown MDP. We depict the progress for robust IMDP policy learning (purple) and robust meta reinforcement learning (yellow). The dotted red line corresponds to the *existential guarantee* [6], i.e., the minimum performance on any MDP from the verification set, when applying the individual optimal policies, which constitutes a natural upper bound on robust policy performance. Figure 6 illustrates the risk-tuning with performance guarantees obtained via Theorem 2. We depict the performances on the PAC IMDPs learned for the verification set (pink dots) and the performance guarantees $\tilde{J}_{(k)}$ when discarding the $k = 0, 5,$ and 10 worst-case samples (dashed lines), corresponding to the risk bounds $\varepsilon_{(k)}$ in Table 2. The full results for all policy learning techniques and benchmarks can be found in [42].

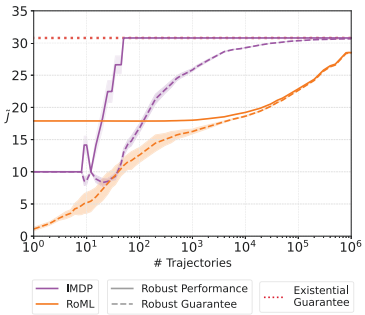


Fig. 5: True robust performances J and guarantees $\tilde{J}$ against number of processed trajectories (betting game).

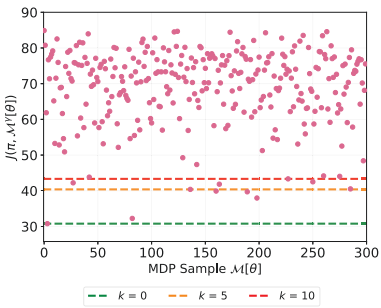


Fig. 6: Robust performance guarantees $\tilde{J}_{(k)}$, when discarding the $k = 0, 5$, and 10 worst-case samples (betting game).

Discussion. The results show that our framework generates tight bounds on the performance and risk of learned policies in upMDPs. Our approach effectively addresses and integrates the two layers of uncertainty: (1) an unknown environment distribution and (2) unknown sample environments. Our results yield high-quality risk bounds for the performance of policies in unseen environments. They enable tuning the risk-performance trade-off, and despite incorporating two layers of uncertainty, provide useful bounds with high user-specified confidence, constituting the first PAC guarantee for this general setup. While policy learning and solving PAC IMDPs scales with the model size and the number of sample MDPs, the computation of the risk bounds via Theorems 1 and 2 depends solely on the number of verification samples N and is independent of the model size. Regarding scalability, we briefly note that the range of model sizes we handle (see Table 1) includes the largest instances handled by comparable methods that perform PAC IMDP learning from trajectories [3,48,33].

5 Conclusion

We have presented a novel approach for producing certifiably robust policies for MDPs with epistemic uncertainty, where transition probabilities depend on parameters with unknown distributions. We have demonstrated that our approach yields tight bounds on a policy's performance in unseen environments from the same distribution. Future work includes extending certifiably robust policies to settings where the specification or task is also uncertain and parameter-dependent.

Acknowledgments. This work was part funded by the ARIA projects SAINT and Super Martingale Certificates, the UKRI AI Hub on Mathematical Foundations of AI and the ERC under the European Union's Horizon 2020 research and innovation programme (FUN2MODEL, grant agreement No. 834115).

References

1. Agha, G., Palmskog, K.: A survey of statistical model checking. ACM Trans. Model. Comput. Simul. **28**(1), 6:1–6:39 (2018)
2. Araya-López, M., Buffet, O., Thomas, V., Charpillet, F.: Active learning of MDP models. In: EWRL. Lecture Notes in Computer Science, vol. 7188, pp. 42–53. Springer (2011)
3. Ashok, P., Kretínský, J., Weininger, M.: PAC statistical model checking for Markov decision processes and stochastic games. In: CAV (1). Lecture Notes in Computer Science, vol. 11561, pp. 497–519. Springer (2019)
4. Bacci, G., Hansen, M., Larsen, K.G.: Model checking constrained Markov reward models with uncertainties. In: QEST. Lecture Notes in Computer Science, vol. 11785, pp. 37–51. Springer (2019)
5. Badings, T., Simão, T.D., Suilen, M., Jansen, N.: Decision-making under uncertainty: Beyond probabilities (2023)
6. Badings, T.S., Cubuktepe, M., Jansen, N., Junges, S., Katoen, J., Topcu, U.: Scenario-based verification of uncertain parametric MDPs. Int. J. Softw. Tools Technol. Transf. **24**(5), 803–819 (2022)
7. Badings, T.S., Jansen, N., Junges, S., Stoelinga, M., Volk, M.: Sampling-based verification of CTMCs with uncertain rates. In: CAV (2). Lecture Notes in Computer Science, vol. 13372, pp. 26–47. Springer (2022)
8. Bäuerle, N., Ott, J.: Markov decision processes with average-value-at-risk criteria. Math. Methods Oper. Res. **74**(3), 361–379 (2011)
9. Beck, J., Vuorio, R., Liu, E.Z., Xiong, Z., Zintgraf, L.M., Finn, C., Whiteson, S.: A survey of meta-reinforcement learning. CoRR **abs/2301.08028** (2023)
10. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities - A Nonasymptotic Theory of Independence. Oxford University Press (2013)
11. Campi, M.C., Garatti, S.: The exact feasibility of randomized solutions of uncertain convex programs. SIAM J. Optim. **19**(3), 1211–1230 (2008)
12. Campi, M.C., Garatti, S.: A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality. J. Optim. Theory Appl. **148**(2), 257–280 (2011)
13. Campi, M.C., Garatti, S.: Introduction to the scenario approach. SIAM (2018)
14. Chi, Z., Liu, Y., Turrini, A., Zhang, L., Jansen, D.N.: A scenario approach for parametric Markov decision processes. In: Principles of Verification (2). Lecture Notes in Computer Science, vol. 15261, pp. 234–266. Springer (2024)
15. Collins, L., Mokhtari, A., Shakkottai, S.: Task-robust model-agnostic meta-learning. In: NeurIPS (2020)
16. Costen, C., Rigter, M., Lacerda, B., Hawes, N.: Planning with hidden parameter polynomial MDPs. Proceedings of the AAAI Conference on Artificial Intelligence **37**(10), 11963–11971 (Jun 2023)
17. Daca, P., Henzinger, T.A., Kretínský, J., Petrov, T.: Faster statistical model checking for unbounded temporal properties. In: TACAS. Lecture Notes in Computer Science, vol. 9636, pp. 112–129. Springer (2016)
18. Derman, E., Mankowitz, D., Mann, T., Mannor, S.: A Bayesian approach to robust reinforcement learning (2019)
19. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML. Proceedings of Machine Learning Research, vol. 70, pp. 1126–1135. PMLR (2017)

20. Ghosh, D., Rahme, J., Kumar, A., Zhang, A., Adams, R.P., Levine, S.: Why generalization in RL is difficult: Epistemic POMDPs and implicit partial observability (2021)
21. Givan, R., Leach, S.M., Dean, T.L.: Bounded-parameter Markov decision processes. Artif. Intell. **122**(1-2), 71–109 (2000)
22. Greenberg, I., Mannor, S., Chechik, G., Meirom, E.A.: Train hard, fight easy: Robust meta reinforcement learning. In: NeurIPS (2023)
23. Gupta, A., Mendonca, R., Liu, Y., Abbeel, P., Levine, S.: Meta-reinforcement learning of structured exploration strategies. In: NeurIPS (2018)
24. Hansson, H., Jonsson, B.: A logic for reasoning about time and reliability. Formal Aspects Comput. **6**(5), 512–535 (1994)
25. Hartmanns, A., Klauck, M., Parker, D., Quatmann, T., Ruijters, E.: The quantitative verification benchmark set. In: TACAS (1). Lecture Notes in Computer Science, vol. 11427, pp. 344–350. Springer (2019)
26. Hérault, T., Lassaigne, R., Magniette, F., Peyronnet, S.: Approximate probabilistic model checking. In: VMCAI. Lecture Notes in Computer Science, vol. 2937, pp. 73–84. Springer (2004)
27. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Springer Series in Statistics (1994)
28. Iyengar, G.N.: Robust dynamic programming. Math. Oper. Res. **30**(2), 257–280 (2005)
29. Junges, S., Jansen, N., Dehnert, C., Topcu, U., Katoen, J.: Safety-constrained reinforcement learning for MDPs. In: TACAS. Lecture Notes in Computer Science, vol. 9636, pp. 130–146. Springer (2016)
30. Kochenderfer, M.: Decision Making Under Uncertainty: Theory and Application (2015)
31. Kwiatkowska, M.Z., Norman, G., Parker, D.: PRISM 4.0: Verification of probabilistic real-time systems. In: CAV. Lecture Notes in Computer Science, vol. 6806, pp. 585–591. Springer (2011)
32. Legay, A., Delahaye, B., Bensalem, S.: Statistical model checking: An overview. In: RV. Lecture Notes in Computer Science, vol. 6418, pp. 122–135. Springer (2010)
33. Meggendorfer, T., Weininger, M., Wienhöft, P.: What are the odds? improving the foundations of statistical model checking. CoRR **abs/2404.05424** (2024)
34. Nilim, A., Ghaoui, L.E.: Robust control of Markov decision processes with uncertain transition matrices. Oper. Res. **53**(5), 780–798 (2005)
35. Pnueli, A.: The temporal logic of programs. In: FOCS. pp. 46–57. IEEE Computer Society (1977)
36. Polgreen, E., Wijesuriya, V.B., Haesaert, S., Abate, A.: Data-efficient Bayesian verification of parametric Markov chains. In: QEST. Lecture Notes in Computer Science, vol. 9826, pp. 35–51. Springer (2016)
37. Poupart, P., Vlassis, N., Hoey, J., Regan, K.: An analytic solution to discrete Bayesian reinforcement learning. In: ICML. ACM International Conference Proceeding Series, vol. 148, pp. 697–704. ACM (2006)
38. Raskin, J., Sankur, O.: Multiple-environment Markov decision processes. In: FSTTCS. LIPIcs, vol. 29, pp. 531–543. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2014)
39. Rickard, L., Abate, A., Margellos, K.: Learning robust policies for uncertain parametric Markov decision processes. In: Proc. L4DC'24. Proceedings of Machine Learning Research, vol. 242, pp. 876–889. PMLR (2024)
40. Ross, S.M.: Introduction to Probability Models. Academic Press (2014)

41. Scheftelowitsch, D., Buchholz, P., Hashemi, V., Hermanns, H.: Multi-objective approaches to Markov decision processes with uncertain transition parameters. In: VALUETOOLS. pp. 44–51. ACM (2017)
42. Schnitzer, Y., Abate, A., Parker, D.: Learning provably robust policies in uncertain parametric environments. CoRR **abs/2408.03093** (2024)
43. Sen, K., Viswanathan, M., Agha, G.: Statistical model checking of black-box probabilistic systems. In: CAV. Lecture Notes in Computer Science, vol. 3114, pp. 202–215. Springer (2004)
44. Steimle, L.N., Kaufman, D.L., Denton, B.T.: Multi-model Markov decision processes. IISE Trans. **53**(10), 1124–1139 (2021)
45. Strehl, A.L., Littman, M.L.: A theoretical analysis of model-based interval estimation. In: ICML. ACM International Conference Proceeding Series, vol. 119, pp. 856–863. ACM (2005)
46. Stückler, J., Schwarz, M., Schadler, M., Topalidou-Kyniazopoulou, A., Behnke, S.: Nimbro explorer: Semiautonomous exploration and mobile manipulation in rough terrain. Journal of Field Robotics (2015)
47. Suilen, M., Badings, T.S., Bovy, E.M., Parker, D., Jansen, N.: Robust markov decision processes: A place where AI and formal methods meet. In: Principles of Verification (3). Lecture Notes in Computer Science, vol. 15262, pp. 126–154. Springer (2024)
48. Suilen, M., Simão, T.D., Parker, D., Jansen, N.: Robust anytime learning of Markov decision processes. In: NeurIPS (2022)
49. Sutton, R.S., McAllester, D.A., Singh, S., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: NIPS. pp. 1057–1063. The MIT Press (1999)
50. Teh, Y.W., Bapst, V., Czarnecki, W.M., Quan, J., Kirkpatrick, J., Hadsell, R., Heess, N., Pascanu, R.: Distral: Robust multitask reinforcement learning. In: NIPS. pp. 4496–4506 (2017)
51. Towers, M., Kwiatkowski, A., Terry, J.K., Balis, J.U., Cola, G.D., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J.J., Tan, H., Younis, O.G.: Gymnasium: A standard interface for reinforcement learning environments. CoRR **abs/2407.17032** (2024)
52. Wang, Y., Zou, S.: Online robust reinforcement learning with model uncertainty (2021)
53. Wiesemann, W., Kuhn, D., Rustem, B.: Robust Markov decision processes. Math. Oper. Res. **38**(1), 153–183 (2013)
54. Wilson, E.B.: Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association (1927)
55. Wolff, E.M., Topcu, U., Murray, R.M.: Robust control of uncertain markov decision processes with temporal logic specifications. In: CDC. IEEE (2012)
56. Younes, H.L.S., Simmons, R.G.: Probabilistic verification of discrete event systems using acceptance sampling. In: CAV. Lecture Notes in Computer Science, vol. 2404, pp. 223–235. Springer (2002)

Open Access. This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

