# Multi-Environment MDPs with Prior and Universal Semantics

**Benjamin Bordais**
Université Libre de Bruxelles, Belgium

**Jean-François Raskin**
Université Libre de Bruxelles, Belgium

## Abstract

Multiple-environment Markov decision processes (MEMDPs) equip an MDP with several probabilistic transition functions (one per possible environment) so that the state is observable but the environment is not. Previous work studies two semantics: (i) the universal semantics, where an adversary picks the environment; and (ii) the prior semantics, where the environment is drawn once before execution from a fixed distribution. We clarify the relation between these semantics. For parity objectives, we show that the qualitative questions, i.e. value one, coincide, and we develop a new algorithm for the general value of MEMDP with prior semantics. In particular, we show that the prior value of an MEMDP with a parity objective can be approximated to any precision with a space efficient algorithm; equivalently, the associated gap problem is decidable in PSPACE when probabilities are given in unary (and in EXPSPACE otherwise). We then prove that the universal value equals the infimum of prior values over all beliefs. This yields a new algorithm for the universal gap problem with the same complexity (PSPACE for unary probabilities, EXPSPACE in general), improving on earlier doubly-exponential-space procedures. Finally, we observe that MEMDPs under the prior semantics form an important tractable subclass of POMDPs: our algorithms exploit the fact that belief entropy never increases, and we establish that any POMDP with this property reduces effectively to a prior-MEMDP, showing that prior-MEMDPs capture a broad and practically relevant subclass of POMDPs.

**2012 ACM Subject Classification** Theory of computation → Logic and verification; Theory of computation → Random walks and Markov chains

**Keywords and phrases** Multi-Environement MDP, approximation algorithm, Partially-observable MDP

**Digital Object Identifier** 10.4230/LIPIcs.ICALP.2026.171

**Category** Track B: Automata, Logic, Semantics, and Theory of Programming

**Related Version** *Full Version*: `https://arxiv.org/abs/2602.10938` [2]

**Funding** The work presented in this article was supported by the FNRS-DFG Weave project "Mixing Formal Methods and Learning Techniques for Strategy Synthesis in Partially Observable Markov Decision Processes."

**Acknowledgements** We also wish to thank Prof. Guillermo Perez and Pierre Vandenhove for their valuable discussions on the topic of this paper.

## 1 Introduction

Multiple-environment Markov decision processes (MEMDPs), first introduced in [13], generalize classical MDPs. They model decision-making problems where the system state is perfectly observable, while the stochastic *environment* is unknown but fixed and chosen or drawn before execution from a finite set of candidates. Formally, an MEMDP consists of a common state and action space together with a probabilistic transition function per possible environment. A single controller acts without being explicitely revealed which environment is

operating. Intuitively, this model lies between fully observable MDPs and partially observable Markov decision processes (POMDPs): it captures a natural class of partial-information problems while avoiding many undecidability barriers of general POMDPs.

Initial work introduced MEMDPs and showed that key synthesis questions that are undecidable for POMDPs become decidable (and sometimes efficiently so) in MEMDPs, thereby justifying the model as a tractable yet expressive *variant* of POMDPs [13]. Subsequent work further developed the theory for $\omega$-regular objectives and clarified the algorithmic landscape [6, 17]. Beyond foundational interest, MEMDPs have been advocated as a practical modeling tool: their structure enables cheaper belief updates than in general POMDPs and supports applications such as contextual recommendation and parameter uncertainty in probabilistic models [4].

### Universal and prior semantics

In previous works, MEMDPs have been studied under two different semantics: [13, 17, 6] study the model under the *universal* semantics and [4] studies the model under the *prior* semantics. In the *universal* (adversarial) semantics, an adversary picks the environment that the fixed controller strategy has to face; the value of a strategy is its probability to satisfy the objective in the worst environment. In the *prior* semantics, the environment is drawn initially at random from a fixed distribution (the prior, known to the player); the value is the corresponding expectation. In both cases, the chosen environment is not revealed to the controller. We focus on parity objectives, that are a canonical way of expressing $\omega$-regular properties, and establish structural and algorithmic links between these semantics.

Before presenting our main contributions, we note a *qualitative* equivalence between the two semantics. The value 1 notions coincide in both the *almost-sure* and *limit-sure* senses: there exist strategies that achieve value 1 (resp. arbitrarily close to 1) in the universal semantics if and only if there exist strategies whose prior value is 1 (resp. arbitrarily close to 1). This correspondence lets us transfer value 1 results between the two semantics (see Corollary 3).

### Contributions

Our main contributions can be summarized as follows:

1. **Approximating the prior value.** We give an algorithm that approximates the *prior* value of an MEMDP to arbitrary precision. This is the most technically demanding part of our contribution. More precisely, we show how to solve algorithmically the $\varepsilon$-gap problem for the prior semantics. This problem asks, given an MEMDP $M$, a prior belief $b$ about the operating environment, a parity objective $W$, a threshold $0 < \alpha < 1$, and a precision $\varepsilon > 0$, to answer YES if the *prior value* is at least $\alpha$, NO if this value is at most $\alpha - \varepsilon$, with no requirement otherwise. Our algorithm runs in PSPACE when probabilities are given in unary, and in EXPSPACE otherwise (see Theorem 10).
2. **From prior to universal.** We also show that the *universal* value equals the infimum, over all prior beliefs, of the corresponding *prior* values (see Theorem 11). Using this result together with the 1-Lipschitz continuity of the prior value with respect to the prior belief, we obtain a new algorithm for the universal $\varepsilon$-gap problem with the same complexity as for the prior value (PSPACE for unary probabilities, EXPSPACE in general), see Theorem 15, improving on earlier doubly-exponential-space procedures proposed in [6].
3. **MEMDPs with prior semantics as a tractable POMDP subclass.** A key reason for MEMDP tractability is that the hidden environment is fixed, which constrains belief dynamics: the expected (information theory) entropy of the belief about the operating

environment is non-increasing. In other words, as a belief entropy measures the amount of information carried by that belief, this means that, on average, the information carried by the environment belief over time does not decrease. We show that in fact any POMDP whose belief entropy is non-increasing can be reduced to a prior-MEMDP (at an exponential cost). Thus, our algorithms apply beyond "pure" MEMDPs and capture a significant subclass of POMDPs, while avoiding classical undecidability phenomena for $\omega$-regular objectives (see Theorem 20).

### Related work

As recalled above, MEMDPs were introduced under the universal semantics by Raskin and Sankur in [13] as a formal model for decision making in scenarios where the environment is fixed yet unknown, and is chosen from a finite set of candidate models. Within this framework, several qualitative $\omega$-regular synthesis problems become decidable, in contrast to the corresponding situation for POMDPs. Subsequent work has refined the qualitative complexity landscape. For almost-sure satisfaction of Rabin objectives, Suilen et al. establish a PSPACE-complete decision procedure [17]. For parity objectives, Chatterjee et al. prove that the value-1 problem is PSPACE-complete in general (and solvable in PTime when the number of environments is fixed), demonstrate that pure strategies suffice to achieve value 1, and present a double-exponential-space approximation scheme for computing the value [6]. Under the prior semantics, MEMDPs have further been proposed as a computationally tractable subclass of POMDPs with discounted payoffs, leveraging more efficient belief-state updates and a non-increasing belief-entropy property [4].

For context, classical (fully observable) MDP qualitative problems are solvable in polynomial time [1], while analogous POMDP questions are much harder as most of them are undecidable [12, 11], and only bounded horizon questions are known to be decidable [12]. So, in this context, MEMDPs occupy a particular place: they retain an interesting part of the partial-information expressiveness but avoid key undecidability barriers that arise in POMDPs. The positive results obtained in this paper add and refine those obtained in [13, 4, 17, 6].

In [7], the authors independently introduced a subclass of POMDPs, coined Posterior-Deterministic POMDPs, which coincides with the Dirac-preserving POMDPs that we define in Section 5.

We additionally refer to [3], which consider *Multi-environment* POMDPs to *both* handle partial observability *and* multiple environments, and studies both finite-horizon and infinite-horizon discounted-reward objectives. Although their framework is more general than ours, they target a different class of objectives, since we focus on undiscounted infinite-horizon objectives. In this setting, the authors propose both exact and approximate solution algorithms.

### Structure of the paper

Section 2 fixes the notation and formally defines MEMDPs together with the universal and prior values. Section 3 addresses the qualitative and quantitative problems for the prior semantics: it establishes the value-1 equivalence with the universal semantics and develops our approximation algorithm for the prior $\varepsilon$-gap problem. Section 4 relates the two semantics by proving that the universal value is the infimum of prior values over all beliefs, and derives an improved complexity for the universal gap problem. Section 5 characterizes MEMDPs as exactly the POMDPs with non-increasing belief entropy, up to an exponential blow-up. Missing details and full proofs can be found in the extended version of the paper [2].

## 2 Definitions

Consider a non-empty set $Q$. The *support* $\mathsf{Supp}(d)$ of a function $d\colon Q \to [0,1]$ is the set $\mathsf{Supp}(d) := \{q \in Q \mid d(q) > 0\}$. A function $d\colon Q \to [0,1]$ is a *distribution* over $Q$ if it has *countable* support and $\sum_{q\in Q} d(q) = 1$. We let $\mathcal{D}(Q)$ denote the set of all probability distributions over the set $Q$. A probability distribution $d \in \mathcal{D}(Q)$ is *Dirac* if $|\mathsf{Supp}(b)| = 1$. For all sets $A$ and $\rho = q_0 \cdot (a_1, q_1) \cdots (a_n, q_n) \in Q \cdot (A \cdot Q)^*$, we let $\mathsf{last}(\rho) := q_n \in Q$.

### MDPs and strategies

A *Markov Decision Process* (MDP for short) $\mathsf{G}$ is a tuple $\mathsf{G} = (Q, A, \delta)$ where $Q$ is a non-empty finite set of *states*, $A$ is a non-empty finite set of *actions*, and $\delta\colon Q \times A \to \mathcal{D}(Q)$ is the transition function mapping each state-action pair to a probability distribution over successor states. An element in $Q \cdot (A \cdot Q)^*$ (resp. $(Q \cdot A)^\omega$) is called a finite (resp. infinite) *run*, an element in $Q^*$ (resp. $Q^\omega$) is called a finite (resp. infinite) *path*. A *strategy* on $(Q, A)$ is a function $\sigma\colon Q \cdot (A \cdot Q)^* \to \mathcal{D}(A)$ mapping finite runs to probability distributions over actions. We let $\mathsf{Strat}(Q, A)$ denote the set of all strategies on $(Q, A)$.

### Objectives

Consider a non-empty finite set $Q$. An *objective* $W \subseteq Q^\omega$ is a Borel set: $W \in \mathsf{Borel}(Q)$. We focus on parity objectives, i.e. objectives $W \in \mathsf{Borel}(Q)$ such that there is a labeling function $f\colon Q \to \mathbb{N}$ such that $W$ is the set of infinite paths whose highest label seen infinitely often is even: $W = \{\rho \in Q^\omega \mid \max\{n \in \mathbb{N} \mid \forall i \in \mathbb{N},\ \exists j \geq i : f(\rho_j) = n\} \text{ is even}\}$.

Consider an MDP $\mathsf{G} = (Q, A, \delta)$. A strategy $\sigma\colon Q \cdot (A \cdot Q)^* \to \mathcal{D}(A)$ induces a probability measure on the set of finite runs, which can be canonically extended to the associated Borel $\sigma$-algebra over infinite runs. We first define the probability measure induced by a strategy after a fixed finite run $\rho \in Q \cdot (A \cdot Q)^*$. For $\pi \in Q \cdot (A \cdot Q)^*$, we define the value $\mathbb{P}^\sigma_\rho(\mathsf{G}, \pi) \in [0,1]$ by induction on the length of $\pi$ as follows. If $|\pi| \leq |\rho|$, then we put $\mathbb{P}^\sigma_\rho(\mathsf{G}, \pi) := 1$ if $\pi$ is a prefix of $\rho$, and $\mathbb{P}^\sigma_\rho(\mathsf{G}, \pi) := 0$ otherwise. For all $\pi \in Q \cdot (A \cdot Q)^*$ such that $\rho$ is a prefix of $\pi$ and for all $(a, t) \in A \times Q$, we define $\mathbb{P}^\sigma_\rho(\mathsf{G}, \pi \cdot (a, t)) := \mathbb{P}^\sigma_\rho(\mathsf{G}, \pi) \cdot \sigma(\pi)(a) \cdot \delta(\mathsf{last}(\pi), a)(t)$.

Then, for all $\pi \in Q \cdot (A \cdot Q)^*$, we define the probability $\mathbb{P}^\sigma_\rho[\mathsf{G}, \mathsf{cyl}(\pi)]$ of the cylinder set $\mathsf{cyl}(\pi) := \{\pi \cdot \rho' \mid \rho' \in (A \cdot Q)^\omega\} \in \mathsf{Borel}(Q \cdot A)$ by $\mathbb{P}^\sigma_\rho[\mathsf{G}, \mathsf{cyl}(\pi)] := \mathbb{P}^\sigma_\rho(\mathsf{G}, \pi) \in [0,1]$. We naturally extend this into a probability measure on Borel sets: $\mathbb{P}^\sigma_\rho[\mathsf{G}, \cdot]\colon \mathsf{Borel}(Q \cdot A) \to [0,1]$. We obtain the probability measure of Borel sets in $\mathsf{Borel}(Q)$ via projection.

### MEMDPs

A *Multi-Environment Markov Decision Process* (MEMDP for short) $\Gamma$ is a tuple $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$ where $E$ is a non-empty finite set of environments, and for all $e \in E$, $(Q, A, \delta_e)$ is an MDP, which we denote $\Gamma[e]$. Unless otherwise stated, an MEMDP $\Gamma$ refers to the tuple $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$. For all $e \neq e' \in E$, a state-action pair $(q, a) \in Q \times A$ is $(e, e')$-*distinguishing* if $\delta_e(q, a) \neq \delta_{e'}(q, a)$. We let $\mathsf{Dstg}(\Gamma, e, e')$ denote the set of all $(e, e')$-distinguishing state-action pairs. We also let $\mathsf{Dstg}(\Gamma) := \cup_{e\neq e'\in E}\mathsf{Dstg}(\Gamma, e, e')$.

### Values in MEMDPs

Consider an MEMDP and a parity objective whose probability we seek to maximize. When synthesizing a strategy in an MEMDP, we do not know in which environment it executes. Thus, when defining the value of a synthesized strategy, we may require either that it

performs well in *all* environments or, given a prior belief on the operating environment, that it performs well on *average*. The goal of this paper is to study the latter "prior" value, and to link it to the former "universal" value. Specifically, given an MEMDP $\Gamma$, a state $q \in Q$, and a parity objective $W \in \mathsf{Borel}(Q)$, we define $\mathsf{uni}$-values

$$\mathsf{val}_q^{\mathsf{uni}}(\Gamma, W) := \sup_{\sigma \in \mathsf{Strat}(Q,A)} \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \sigma, W) \text{ with } \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \sigma, W) := \min_{e \in E} \mathbb{P}_q^{\sigma}[\Gamma[e], W]$$

and $\mathsf{pr}$-values, given a prior belief $b \in \mathcal{D}(E)$

$$\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) := \sup_{\sigma \in \mathsf{Strat}(Q,A)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma, W) \text{ with } \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma, W) := \sum_{e \in E} b(e) \cdot \mathbb{P}_q^{\sigma}[\Gamma[e], W]$$

▶ **Example 1.** Let us consider Fig. 1, which represents an MEMDP. States, actions, and probabilistic successors are defined as in a classical MDP, with the key difference that here we consider multiple valuations of the parameters that label the probabilistic transitions (i.e., the parameters on the outgoing edges of the triangular nodes).

As a first example, assume that we want to model a deck of cards in which card 1 has two copies and card 2 has one copy. In this case, the probability of drawing card 1 is $\alpha_1 = \frac{2}{3}$, while the probability of drawing card 2 is $\alpha_2 = \frac{1}{3}$. Suppose moreover that the player is always allowed to request another draw before making a guess; this is captured by setting $\alpha_0 = 0$.

To model the fact that the duplicated card is card 1, we assign the parameters $\beta_1 = 1$ and $\beta_2 = 0$, meaning that guessing card 1 leads with certainty to the winning state $W$, while guessing card 2 leads to the losing state $L$. Symmetrically, we set $\gamma_1 = 0$ and $\gamma_2 = 1$. We refer to this valuation of the parameters as environment $E_1$. We can define another environment, denoted $E_2$, that models the situation in which card 2 is duplicated instead. In this case, the parameters are set to $\alpha_1 = \frac{1}{3}$ and $\alpha_2 = \frac{2}{3}$, together with $\beta_1 = 0$, $\beta_2 = 1$, $\gamma_1 = 1$, and $\gamma_2 = 0$.

Finally, to encode the reachability objective (i.e., reaching state $W$), we use the following parity labeling: all states have label 1, except the winning state $W$, which has label 2. Under this labelling, the parity objective is satisfied if and only if the play eventually reaches $W$, i.e., the player correctly guesses the duplicated card.

## 3 Qualitative and quantitative problem

In this section, we focus on deciding the existence of strategies with good enough prior-values. We first study the qualitative problems (value 1) and then turn to the quantitative problem.

### 3.1 Qualitative problem

For value (arbitrarily close to) 1, the universal and prior settings coincide.

▶ **Proposition 2.** *Consider an MEMDP $\Gamma$, a state $q \in Q$, and a parity objective $W$. Let $b \in \mathcal{D}(E)$ such that $\mathsf{Supp}(b) = E$. Then $\mathsf{val}_q^{\mathsf{uni}}(\Gamma, W) = 1$ if and only if $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) = 1$ and, for all $\sigma \in \mathsf{Strat}(Q, A)$: $\mathsf{val}_q^{\mathsf{uni}}(\Gamma, \sigma, W) = 1$ if and only if $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma, W) = 1$.*

**Proof sketch.** If $\mathsf{val}_q^{\mathsf{uni}}(\Gamma, W) = 1$, then there are strategies whose $\mathsf{uni}$-values are arbitrarily close to 1, thus their $\mathsf{pr}$-values are arbitrarily close to 1. Hence, $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, W) = 1$. Furthermore, if a strategy has a $\mathsf{pr}$-value, w.r.t. $b$, of at least $1 - d \cdot \varepsilon$ for some $\varepsilon > 0$, with $d := \min_{e \in E} b(e) > 0$, then its smallest value in any environment (since $\mathsf{Supp}(b) = E$) is at least $1 - \varepsilon$. Hence, if $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, W) = 1$ then $\mathsf{val}_q^{\mathsf{uni}}(\Gamma, W) = 1$. The other equivalence is straightforward. ◀

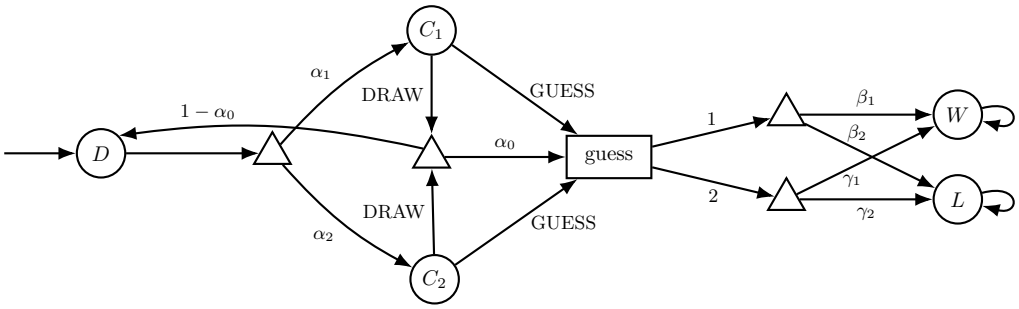


**Figure 1** The figure depicts an MEMDP (inspired from [6]) modelling a simple card game played with a deck containing two card types, 1 and 2. The deck composition is parameterized by $\alpha_1$ and $\alpha_2$, with $\alpha_1 + \alpha_2 = 1$; for instance, if card 1 appears twice as often as card 2, then $\alpha_1 = 2\alpha_2$. At each turn, the player may either *guess* which card type is in the majority or request an additional *draw* from the deck; after a draw request, the game continues with probability $1 - \alpha_0$, while with probability $\alpha_0$ the player is forced to guess immediately. The player's objective is to reach the winning state $W$, which corresponds to correctly guessing the card type that has the larger number of copies in the deck; this is captured by suitable choices of the parameters (e.g., $\beta$ and $\gamma$) governing the transition from the guess action to $W$ or $L$.

The complexity of deciding the existence of almost-surely winning strategies (value 1) and of limit-surely winning strategies (values arbitrarily close to 1) is thus the same in MEMDPs with parity objectives in the universal and prior settings. We deduce the corollary below.

▶ **Corollary 3** (of [6, Theorems 4, 13]). *In an MEMDP, given a prior belief and a parity objective, the problem of deciding the existence of strategies of* pr*-value (arbitrarily close to) 1 is* PSPACE*-complete. When the number of environments is fixed, it can be solved in polynomial time.*

▶ **Example 4.** Consider again the MEMDP from Example 1. It is easy to see that no strategy can guarantee winning with probability 1. Indeed, after any finite number of draws, the empirical frequencies may still be atypical and hence suggest the wrong duplicated card.

Nevertheless, for every $\varepsilon > 0$, the player can request sufficiently many draws so that the probability of observing such a misleading sample drops below $\varepsilon$. By the law of large numbers, this implies that although almost-sure winning is impossible, the player can make the winning probability arbitrarily close to 1 by choosing an appropriate family of strategies (indexed by $\varepsilon$).

This reasoning applies both under the universal semantics and under the prior semantics whenever the prior distribution assigns positive probability to both environments. In either case there is no surely winning strategy, but there exists a family of strategies witnessing limit-sure winning; in particular, the construction works for any fixed prior over the two environments.

## 3.2 Quantitative problem

Let us now turn to the quantitative problem. We design an algorithm that, for any prescribed precision level $\varepsilon > 0$, computes an $\varepsilon$-approximation of the pr-value $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W)$. This approximation procedure is subsequently employed to solve the associated *gap problem*, defined as follows: given a parameter $\varepsilon > 0$, one must output YES if $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) \geq \alpha$, NO if $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) \leq \alpha - \varepsilon$, and an arbitrary answer otherwise. Since there are inputs for which no specific output is required, the gap problem is not a classical decision problem; it is instead a promise problem [9].

Our goal is to compute an approximation of the pr-value given a prior environment belief $b \in \mathcal{D}(E)$. When playing in an MEMDP, the environment belief is updated each time distinguishing state-action pairs are encountered to account for the change of likelihood of different environments, e.g. when we visit a transition $(q, a, q') \in Q \times A \times Q$ such that $\delta_e(q, a)(q') > \delta_{e'}(q, a)(q')$ for some $e \neq e' \in E$, the environment $e$ is more likely than before compared to environment $e'$ . We formally define below this belief update.

▶ **Definition 5** (Belief update). *Consider an MEMDP $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$. Let $b \in \mathcal{D}(E)$. For all $(q, a) \in Q \times A$, we let $p[b, q, a] \in \mathcal{D}(Q)$ be defined by, for all $q' \in Q$:*

$$p[b, q, a](q') := \sum_{e\in E} b(e) \cdot \delta_e(q, a)(q')$$

*Then, for all $q' \in Q$, we let $\lambda[b, q, a, q'] \in \mathcal{D}(E)$ be an arbitrary distribution if $p[b, q, a](q') = 0$, and otherwise, for all $e \in E$:*

$$\lambda[b, q, a, q'](e) := \frac{b(e) \cdot \delta_e(q, a)(q')}{p[b, q, a](q')}$$

*Given any initial state $q \in Q$ and prior belief $b \in \mathcal{D}(E)$, we define a ("likelihood") function $\mathsf{lk}_b : Q \cdot (A \cdot Q)^* \to \mathcal{D}(E)$ which iteratively updates the environment belief: for all $q \in Q$, $\mathsf{lk}_b(q) := b \in \mathcal{D}(E)$ and, for all $\rho \cdot (a, q') \in Q \cdot (A \cdot Q)^+$, we let $\mathsf{lk}_b(\rho \cdot (a, q')) := \lambda[\mathsf{lk}_b(\rho), \mathsf{last}(\rho), a, q']$.*

▶ **Example 6.** Let us illustrate how belief updates work on our running example. Assume that the prior over the two environments is given by the distribution $b = (\frac{1}{4}, \frac{3}{4})$, so the initial belief is skewed towards environment $E_2$ (i.e., towards card 2 being duplicated). Now suppose that after the first draw we observe card 2, meaning that we reach state 2. This observation should increase our confidence that the operating environment is indeed $E_2$. Applying the update rule above, the posterior belief becomes $\lambda[b, D, a, C_2](E_2) = \frac{3}{4} \cdot \frac{\frac{2}{3}}{\frac{1}{3}\cdot\frac{1}{4}+\frac{2}{3}\cdot\frac{3}{4}} = \frac{3}{4} \cdot \frac{24}{21} = \frac{6}{7}$, and $\lambda[b, D, a, C_2](E_1) = \frac{1}{7}$. Conversely, if the first draw yields card 1, then the posterior belief becomes $\lambda[b, D, a, C_1](E_2) = \frac{3}{4} \cdot \frac{\frac{1}{3}}{\frac{2}{3}\cdot\frac{1}{4}+\frac{1}{3}\cdot\frac{3}{4}} = \frac{3}{4} \cdot \frac{12}{15} = \frac{3}{5}$, and $\lambda[b, D, a, C_1](E_1) = \frac{2}{5}$.

Our algorithm computing an approximation of the pr-value works recursively on the number of environments in the support of the belief. When there is a single environment in the support, we are actually in an MDP, and we can compute in polynomial time the values of all states [1]. Assume now that the support of the prior belief is of size at least 2. If the belief in an environment $e$ ever drops to 0, because we have visited a transition $(q, a, q')$ such that $\delta_e(q, a)(q') = 0$, we can call the algorithm recursively with a belief of smaller support. Alternatively, the belief in an environment may be close to 0. In that case, it is almost harmless (in terms of value change) to truncate the belief into one with a smaller support, by 1-Lipschitz continuity of the pr-value, stated in the lemma below.

▶ **Lemma 7.** *Consider an MEMDP $\Gamma$ and a parity objective $W$. For all beliefs $b, b' \in \mathcal{D}(E)$, we have $|\mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, W) - \mathsf{val}^{\mathsf{pr}}_q(\Gamma, b', W)| \leq \mathsf{Diff}(b, b')$, with $\mathsf{Diff}(b, b') := \frac{1}{2} \cdot \sum_{e\in E} |b(e) - b'(e)|$. Furthermore, for all beliefs $b \in \mathcal{D}(E)$ such that $|\mathsf{Supp}(b)| \geq 2$, there is $\mathsf{Truncate}(b) \in \mathcal{D}(E)$ such that $\mathsf{Supp}(\mathsf{Truncate}(b)) \subsetneq \mathsf{Supp}(b)$ and $\mathsf{Diff}(b, \mathsf{Truncate}(b)) \leq \min_{e\in\mathsf{Supp}(b)} b(e)$.*

The environment belief is updated each time a distinguishing state-action pair is visited. However, even if many distinguishing state-action pairs are visited, it could be that there is no environment whose belief drops to 0. Nonetheless, if enough distinguishing state-action pairs are visited, with high probability, there is an environment whose belief drops close to 0.

▶ **Theorem 8.** *Consider an MEMDP* $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$ *and a prior environment belief* $b \in \mathcal{D}(E)$*. Let* $k := |E|$*, and* $l$ *denote the maximum number of bits to write any probability occurring in* $\Gamma$*. For all* $\varepsilon > 0$*, we let:* $\mathsf{NeverSmallBelief}(b, \varepsilon) := \{\rho \in Q \cdot (A \cdot Q)^* \mid \forall \pi \sqsubseteq \rho,\ \forall e \in E :\ \mathsf{lk}_b(\pi)(e) > \varepsilon\}$*, with* $\pi \sqsubseteq \rho$ *meaning that* $\pi$ *is a prefix of* $\rho$*. For all* $\rho \in (Q \cdot A)^\omega$*, we let* $\mathsf{nb}_{\mathsf{dstg}}(\rho) \in \mathbb{N} \cup \{\infty\}$ *denote the number of times* $\rho$ *visits a distinguishing state-action pair.*

*Then, for all* $\varepsilon > 0$*, letting* $x := \lceil \log(\frac{1}{\varepsilon}) \rceil$*, there is* $m(\Gamma, \varepsilon) \in \mathbb{N}$ *such that* $m(\Gamma, \varepsilon) = O(poly(k, 2^l, x))$ *ensuring that for all strategies* $\sigma \in \mathsf{Strat}(Q, A)$ *and environments* $e \in E$*:*

$$\mathbb{P}^\sigma_q(\Gamma[e], \{\mathsf{nb}_{\mathsf{dstg}} \geq m(\Gamma, \varepsilon)\} \cap \mathsf{NeverSmallBelief}(b, \varepsilon)) \leq \varepsilon$$

This theorem constitutes a crucial result of this paper, its proof is rather intricate. Below, we only describe the two main ingredients of that proof.

**Proof sketch.** Assume that some $e \in \mathsf{Supp}(b)$ is the operating environment. Consider another environment $e' \neq e \in \mathsf{Supp}(b)$ and assume, for simplicity, that $\mathsf{Supp}(\delta_e(q, a)) = \mathsf{Supp}(\delta_{e'}(q, a))$ for all $(q, a) \in Q \times A$. Our goal is to show that, regardless of the strategy $\sigma$, with high probability, runs $\rho$ that visit enough $(e, e')$-distinguishing state-action pairs are such that $\mathsf{lk}_b(\rho)(e')$ is arbitrarily small. We focus on the ratio of the beliefs in the environments $e$ and $e'$, which initially is equal to $\frac{b(e')}{b(e)}$. After some run $\rho \in Q \cdot (A \cdot Q)^*$, it is equal to $\frac{\mathsf{lk}_b(\rho)(e')}{\mathsf{lk}_b(\rho)(e)} = \frac{b(e')}{b(e)} \cdot \prod_{\pi\cdot(a,q)\sqsubseteq\rho} \frac{\delta_{e'}(\mathsf{last}(\pi),a)(q)}{\delta_e(\mathsf{last}(\pi),a)(q)}$. Since the sum of random variables is easier to analyze than their product, we aim at showing $(*)$: "Regardless of the strategy $\sigma$, with high probability, if $\rho$ visits enough $(e, e')$-distinguishing state-action pairs, the sum $\sum_{\pi\cdot(a,q)\sqsubseteq\rho} \log(\frac{\delta_{e'}(\mathsf{last}(\pi),a)(q)}{\delta_e(\mathsf{last}(\pi),a)(q)})$ is low enough". An instrumental result in the proof of this fact is that there is some $\eta < 0$ such that, for all $(q, a) \in \mathsf{Dstg}(\Gamma, e, e')$:$\sum_{q'\in Q} \delta_e(q, a)(q') \cdot \log(\frac{\delta_{e'}(q,a)(q')}{\delta_e(q,a)(q')}) \leq \eta$. This is a consequence of the difference between the arithmetic and geometric means.

In order to show $(*)$, we want to use Hoeffding's inequality [10] (which was also a central tool used in [6]). Informally, this inequality entails that if we consider $n \in \mathbb{N}$ independent real-valued random variables, the probability that their sum is higher than their expected value decreases exponentially with $n$. In our case, we could consider the random variables $X_k$, for $k \in \mathbb{N}$, that map the $k$-th transition $(q, a, q') \in Q \times A \times Q$ visited for which $(q, a) \in \mathsf{Dstg}(\Gamma, e, e')$ to the real value $\log(\frac{\delta_{e'}(q,a)(q')}{\delta_e(q,a)(q')}) \in \mathbb{R}$. However, these random variables are clearly not independent and thus Hoeffding's inequality cannot be applied. Nonetheless, since we can show that the expected value of all of these random variables is at most $\eta < 0$, we can adapt the proof of Hoeffding's inequality to show that, regardless of the strategy $\sigma$, with probability exponentially small in $n$, the sum of the $n$ first random variables $X_1, \ldots, X_n$ is not much higher than $n \cdot \eta < 0$.

Overall, if we appropriately choose $n$, we obtain that, in the environment $e$, for all strategies $\sigma$, the probability to visit at least $n$ $(e, e')$-distinguishing state-action pairs while having the belief in the environment $e'$ never dropping below $\varepsilon$ to be at most $\varepsilon > 0$. The quantity $m(\Gamma, \varepsilon)$ is chosen such that if at least $m(\Gamma, \varepsilon)$ distinguishing state-action pairs are visited, for all environments $e \in E$, there is an environment $e' \in E$ such that at least $n$ $(e, e')$-distinguishing state-action pairs are visited. The theorem follows. ◀

▶ **Example 9.** Let us consider our running example with environments $E_1$ and $E_2$ as above, except that we consider a small $\alpha_0 > 0$, so that with high probability, the strategy can gather many observations before having to make a guess. As we have seen in the above example on the update of the belief, visiting state $C_2$ increases the belief in environment $E_2$, and visiting state $C_1$ increases the belief in environment $E_1$. A consequence of the above theorem is that,

**Algorithm 1** MEMDP-Prior-Parity$(\Gamma, W, b, n, \gamma)$.

**Input**: MEMDP $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$, parity objective $W$, belief $b \in \mathcal{D}(E)$, $n \in \mathbb{N}$, $\gamma > 0$

1: **if** $|\mathsf{Supp}(b)| = 1$ **then return** MDP-Parity$(\Gamma, W) : Q \to [0, 1]$
2: **if** $n = 0$ or $\min_{e\in\mathsf{Supp}(b)} b(e) \leq \frac{\gamma}{3|E|}$ **then**
3: $\quad b' \leftarrow \mathsf{Truncate}(b)$ ▷ As in Lemma 7: $\mathsf{Supp}(b') \subsetneq \mathsf{Supp}(b)$
4: $\quad$ **return** MEMDP-Prior-Parity$(\Gamma, W, b', m(\Gamma, \frac{\gamma}{3|E|}), \gamma)$
5: **for** $(q, a) \in \mathsf{Dstg}(\Gamma)$ and $q' \in \mathsf{Supp}(p[b, q, a])$ **do**
6: $\quad b' \leftarrow \lambda[b, q, a, q']$ ▷ As in Definition 5
7: $\quad$ **if** $\mathsf{Supp}(b') \subsetneq \mathsf{Supp}(b)$ **then** $n' \leftarrow \frac{\gamma}{3|E|}$ **else** $n' \leftarrow n - 1$
8: $\quad v_{q,a,q'} \leftarrow$ MEMDP-Prior-Parity$(\Gamma, W, b', n', m)(q')$
9: $(\mathsf{G}, W') \leftarrow$ MDP-from-MEMDP$(\Gamma, W, b, v)$
10: **return** MDP-Parity$(\mathsf{G}, W') : Q \to [0, 1]$

with high probability, the number of visits to the states revealing a card is massively skewed towards either $C_1$ (and the belief in $E_2$ drops close to 0) or $C_2$ (and the belief in $E_1$ drops close to 0). Once this occurs, we can almost harmlessly truncate the belief (recall Lemma 7) and obtain a regular MDP, in which we can compute the value in polynomial time.

### Algorithm solving the gap problem

We have designed Algorithm 1 that, given an MEMDP $\Gamma$, a parity objective $W$, a prior environment belief $b \in \mathcal{D}(E)$ and some $\gamma > 0$, computes, for every state $q \in Q$, a value $v_q$ such that $|\mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, W) - v_q| \leq \gamma$. The algorithm takes as additional argument an integer $n \in \mathbb{N}$ which bounds the maximum number of distinguishing state-action pairs that can be visited before the environment belief is truncated. The algorithm should be called with $n := m(\Gamma, \frac{\gamma}{3|E|})$[1]. This algorithm works recursively on the number of environments in the support of the belief; the base case is handled in Line 1: in an MDP with parity objectives, we can compute in polynomial time the values of all states. Then, if we have visited enough distinguishing state-action pairs ($n = 0$) or the smallest positive belief is sufficiently close to 0, we recursively call the algorithm on a truncated environment belief with smaller support (Lines 2-4). Otherwise, for each distinguishing state-action pair $(q, a)$, and states $q' \in Q$ such that $p[b, q, a](q') > 0$ (recall Definition 5), we compute the value $v_{q,a,q'} \in [0, 1]$ of that state $q'$ with an updated belief $\lambda[b, q, a, q'] \in \mathcal{D}(E)$ by recursively calling the algorithm with a bound $n'$ that is either reset to $m(\Gamma, \frac{\gamma}{3|E|})$ if the support of the belief has shrunk, or equal to $n - 1$ otherwise (Lines 5-8). We then transform the MEMDP into an MDP by adding two fresh sink states $q_{\mathsf{win}}$ and $q_{\mathsf{lose}}$ redirecting all transitions $(q, a, q') \in Q \times A \times Q$ such that $(q, a)$ is a distinguishing state-action pair leading to the state $q_{\mathsf{win}}$ with probability $v_{q,a,q'}$ and to the state $q_{\mathsf{lose}}$ with probability $1 - v_{q,a,q'}$; we also modify the parity objective $W$ into a parity objective $W'$ for which looping on $q_{\mathsf{win}}$ is winning and looping on $q_{\mathsf{lose}}$ is losing (Line 9). We can finally compute the value of the states in that MDP with the objective $W'$ (Line 10). They correspond to the values of the MEMDP $\Gamma$.

Consider now the complexity of Algorithm 1. First, note that, whenever it recursively calls itself, either the support of the belief has shrunk, or the bound (which resets to $m(\Gamma, \frac{\gamma}{3|E|})$) on the number of visited distinguishing state-action pairs has decreased. Thus, the recursion

[1] We use $\frac{\gamma}{3|E|}$ instead of $\gamma$ because the algorithm performs several approximations by truncating environment beliefs, the sum of all these approximations should total at most $\gamma$.

depth is bounded by $|E| \times m(\Gamma, \frac{\gamma}{3|E|})$. In addition, the number of bits used to describe the environment beliefs grows linearly with the number of updates and truncations. The space taken by the algorithm is in fact in $O(poly(|Q|, |A|, |E|, |\log(\gamma)|, x_b, m(\Gamma, \frac{\gamma}{3|E|})))$, where $x_b$ is the number of bits to represent the prior belief $b$.

The gap problem can be solved by calling Algorithm 1 with $\gamma := \varepsilon/3$, and comparing the result with $\alpha$. Given the bound on $m(\Gamma, \frac{\gamma}{3|E|})$ from Theorem 8, we obtain the theorem below.

▶ **Theorem 10.** *In an MEMDP $\Gamma$, given a prior belief $b \in \mathcal{D}(E)$ (given in binary), a parity objective, a threshold $0 < \alpha < 1$ (given in binary), and a precision $\varepsilon > 0$ (given in binary), the gap problem with* pr*-values can be decided in* EXPSPACE*. If the probabilities involved in the MEMDP are given in unary, the gap problem can be decided in* PSPACE*.*

## 4 Relating the pr- and uni-values

The procedure described in [6] that solves the gap problem with uni-values executes in space doubly exponential in $|Q|$ (the number of states). On the other hand, we have exhibited an algorithm solving the gap problem with pr-values in polynomial space when the probabilities are given in unary, in exponential space otherwise. Our goal now is to link the pr- and uni-values so that we can use our algorithm to solve more efficiently the gap problem with uni-values.

It is clear that the uni-value is lower than or equal to the pr-value for any prior belief. In fact, the uni-value is actually equal to the infimum pr-value over all possible prior beliefs.

▶ **Theorem 11.** *Consider an MEMDP $\Gamma = (Q, A, E, (\delta_e)_{e \in E})$, a state $q \in Q$, and a parity objective $W \subseteq \mathsf{Borel}(Q)$. We have: $\mathsf{val}^{\mathsf{uni}}_q(\Gamma, W) = \inf_{b \in \mathcal{D}(E)} \mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, W)$.*

▶ **Example 12.** We reconsider our running example and modify the two environments $E_1$ and $E_2$ as follows. In both environments, we set $\alpha_0 = 1$, so the player is required to guess the duplicated card immediately after a single draw. If all other aspects of the environments remain unchanged, the uni-value is $\frac{2}{3}$; it is attained by the strategy that guesses the environment corresponding to the single observed card. Under a uniform prior over the two environments, the pr-value coincides with this $\frac{2}{3}$ value. We now introduce asymmetric environments. In $E_1$, card 1 is duplicated twice and card 2 is duplicated once (hence $\alpha_1 = \frac{3}{5}$), while in $E_2$, card 2 is duplicated twice (hence $\alpha_1 = \frac{1}{4}$). In this setting, the uni-value remains equal to $\frac{2}{3}$; however, the infimum of the pr-values is now achieved under a prior $b$ that is slightly biased towards the less favorable environment, namely $b(E_1) := \frac{5}{9}$, rather than under the uniform prior over environments.

To establish Theorem 11, we introduce the notion of mixed strategies. Given a set of states $Q$ and set of actions $A$, a mixed strategy on $(Q, A)$ is a probability distribution $\tau \in \mathcal{D}(\mathsf{Strat}(Q, A))$ over the strategies (even though the set $\mathsf{Strat}(Q, A)$ is uncountable, the probability distributions that we consider have a countable support). We naturally define, in an MDP $\mathsf{G} = (Q, A, \delta)$, the probability measure $\mathbb{P}^\tau_\rho[\mathsf{G}, \cdot] : \mathsf{Borel}(Q) \to [0, 1]$ induced by a mixed strategy $\tau \in \mathcal{D}(\mathsf{Strat}(Q, A))$: $\mathbb{P}^\tau_\rho[\mathsf{G}, \cdot] := \sum_{\sigma \in \mathsf{Strat}(Q,A)} \tau(\sigma) \cdot \mathbb{P}^\sigma_\rho[\mathsf{G}, \cdot]$. For all Borel objectives $W \in \mathsf{Borel}(Q)$, we naturally extend to mixed strategies $\tau \in \mathcal{D}(\mathsf{Strat}(Q, A))$ the uni-values $\mathsf{val}^{\mathsf{uni}}_q(\Gamma, \tau, W) \in [0, 1]$ and pr-values $\mathsf{val}^{\mathsf{uni}}_q(\Gamma, b, \tau, W) \in [0, 1]$ given a prior belief $b \in \mathcal{D}(E)$. Then, we can relate the supremum uni-values that mixed strategies can achieve with the infimum pr-values over all prior beliefs.

▶ **Lemma 13.** *Consider an MEMDP $\Gamma$, a state $q \in Q$, and a parity objective $W$. We have:*

$$\sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \inf_{b \in \mathcal{D}(E)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W) = \inf_{b \in \mathcal{D}(E)} \sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W)$$

*Therefore:* $\sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W) = \inf_{b \in \mathcal{D}(E)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W)$.

**Proof sktech.** The first equality is a direct consequence of a standard generalization of von Neuman's minimax theorem [18], which holds because the set of environments is finite [16]. Furthermore, for all non-empty sets $X$ and $f\colon X \to [0,1]$, we have $\sup_{d \in \mathcal{D}(X)} \sum_{x \in X} d(x) \cdot f(x) = \sup_{x \in X} f(x)$ and $\inf_{d \in \mathcal{D}(X)} \sum_{x \in X} d(x) \cdot f(x) = \inf_{x \in X} f(x)$. Therefore, we have that for all $\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))$, $\inf_{b \in \mathcal{D}(E)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W) = \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W)$; and for all $b \in \mathcal{D}(E)$, $\sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W) = \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W)$. ◀

To establish Theorem 11, it is now sufficient to show that mixed strategies in $\mathcal{D}(\mathsf{Strat}(Q,A))$ do not achieve higher $\mathsf{uni}$-value than strategies in $\mathsf{Strat}(Q,A)$, as stated in the lemma below.

▶ **Lemma 14.** *Consider an MEMDP $\Gamma$, a state $q \in Q$, and a parity objective $W \in \mathsf{Borel}(Q)$. We have:* $\sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W) = \mathsf{val}_q^{\mathsf{uni}}(\Gamma, W)$.

**Proof sketch.** Given any $\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))$, we define $\sigma \in \mathsf{Strat}(Q,A)$ such that, for all $e \in E$ and Borel objectives $W \in \mathsf{Borel}(Q)$: $\mathbb{P}_q^{\tau}[\Gamma[e], W] = \mathbb{P}_q^{\sigma}[\Gamma[e], W]$. Thus, $\mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W) = \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \sigma, W) \leq \mathsf{val}_q^{\mathsf{uni}}(\Gamma, W)$. The lemma follows. ◀

Note that a similar result is established in [3, Theorem 2], although with a different proof technique and in a different context (i.e. finite-horizon or infinite-horizon discounted reward instead of parity), as discussed in the introduction.

### How to use Theorem 11

Lemmas 13 and 14 together imply Theorem 11. This theorem gives us that computing the $\mathsf{uni}$-value amounts to computing the infimum of $\mathsf{pr}$-values over all prior beliefs. Furthermore, Lemma 7 gives us that $\mathsf{pr}$-values induced by two close beliefs are not far-off. Therefore, we can obtain an $\varepsilon$-approximation of the $\mathsf{uni}$-value by computing the $\mathsf{pr}$-value for enough prior beliefs that tightly cover the set of all beliefs. We deduce the theorem below that significantly improves the doubly-exponential-space complexity established in [6].

▶ **Theorem 15.** *In an MEMDP given a parity objective, a threshold $0 < \alpha < 1$ (in binary), and a precision $\varepsilon > 0$ (in binary), the gap problem with $\mathsf{uni}$-values can be decided in* EXPSPACE*. If the probabilities are given in unary, the gap problem can be decided in* PSPACE*.*

**Proof sketch.** Let $N := \lceil \log(\frac{1}{\varepsilon}) \rceil$, $k := |E|$, and $S := \{b \in \mathcal{D}(E) \mid \forall e \in E,\ b(e) \in \{\frac{x}{k \cdot 2^{N+1}} \mid 0 \leq x \leq k \cdot 2^{N+1}\}\}$. We have $|S| = O(2^{k^2 \cdot (N+1)})$ and, for all $b \in \mathcal{D}(E)$, there is $b' \in S$ such that $\mathsf{Diff}(b, b') \leq \frac{\varepsilon}{2}$. Enumerating all beliefs in $S$ can be done in space polynomial in $k$ and $N$. Furthermore, all the beliefs in $S$ are described with a number of bits polynomial in $k$ and $N$. Therefore, executing Algorithm 1 on any belief in $S$ with $\gamma := \varepsilon/2$ takes space exponential in the input (resp. polynomial in the input, if the probabilities are given in unary). ◀

Our goal is now to use Theorem 11 to derive a lower bound on the complexity of approximating the $\mathsf{pr}$-value with probabilities written in unary, we have already established a polynomial space upper bound. It was shown in [14, Theorem 26] that the gap problem with $\mathsf{uni}$-values and reachability objectives in two-environment MEMDPs with probabilities

written in unary is NP-hard. Thus, we focus below on how to approximate uni-values in MEMDPs with two-environments. As argued below, this can be done with only polynomially many calls to an oracle approximating the pr-value, which allows the transfer of the above NP-hardness result.

▶ **Proposition 16.** *In an MEMDP $\Gamma$ with two environments, given a parity objective $W$, and a precision $\varepsilon > 0$, letting $N := \lceil \log(\frac{1}{\varepsilon}) \rceil$, we can compute a $\varepsilon$-approximation of the* uni*-value in time polynomial in $N$, with $O(N)$ calls to an oracle computing $\gamma$-approximation of the* pr*-values on beliefs and $\gamma$ described with a number of bits polynomial in $N$.*

**Proof sketch.** Let $E = \{e_1, e_2\}$. We consider $f\colon [0,1] \to [0,1]$ such that, for all $x \in [0,1]$, $f(x) := \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b_x, W)$, with $b_x \in \mathcal{D}(E)$ such that $b_x(e_1) := x$. By Theorem 11, the infimum $f$-value is the uni-value, i.e. $\inf_{x\in[0,1]} f(x) = \mathsf{val}_q^{\mathsf{uni}}(\Gamma, W)$. The function $f$ is quasi-convex: for all $x < y < z \in [0,1]$: $f(y) \leq \max(f(x), f(z))$. This implies that, for all $x < y < z < t$, if $f(z) < f(y)$, then $\inf_{u\in[x,t]} f(u) = \inf_{u\in[y,t]} f(u)$, and if $f(y) < f(z)$, $\inf_{u\in[x,t]} f(u) = \inf_{u\in[x,z]} f(u)$. We design a binary-search-like procedure based on this observation that finds an approximation of the minimal $f$-value by searching in intervals $[x,t]$ of decreasing length (it is multiplied by $\frac{2}{3}$ at each step) until that length becomes small enough (which is sufficient because $f$ is 1-Lipschitz continuous). ◀

▶ **Corollary 17.** *In two-environment MEMDPs with probabilities in unary, deciding the* pr*-value gap problem with parity objectives cannot be done in polynomial time unless* P = NP.

## 5 Characterization of MEMDPs with entropy

When playing in an MEMDP, we have only partial information about where we are: we know the current state, but we do not know the operating environment. In fact, MEMDPs are a special kind of Partially Observable MDPs (POMDP for short), i.e. MDPs in which we play on an underlying set of states to which we have only indirect access via an observation that may be identical for different states. Formally, a POMDP is a tuple $\Lambda = (Q, A, \delta, \Omega, O)$ where $(Q, A, \delta)$ is an MDP, $\Omega$ is a non-empty set of observations, and $O\colon Q \to \Omega$. An MEMDP $\Gamma$ naturally induces the POMDP $\Lambda(\Gamma) = (Q', A, \delta, \Omega, O)$ such that:

- $Q' := Q \times E$;
- for all $(q, e, q', a) \in Q \times E \times Q \times A$, $\delta((q,e),a)((q',e)) := \delta_e(q,a)(q')$;
- $\Omega := Q$; and for all $(q,e) \in Q \times E$, $O((q,e)) := q$.

Strategies in POMDPs are functions in $\mathsf{Strat}(\Omega, A)$. Given any strategy $\sigma \in \mathsf{Strat}(\Omega, A)$, for any state $q \in Q$, we naturally define the probability measure $\mathbb{P}_q^\sigma[\Lambda, \cdot]\colon \mathsf{Borel}(Q) \to [0,1]$. Given $b \in \mathcal{D}(Q)$ and $W \in \mathsf{Borel}(Q)$, we let $\mathsf{val}(\Lambda, b, W) := \sup_{\sigma\in\mathsf{Strat}(O,A)} \sum_{q\in Q} b(q) \cdot \mathbb{P}_q^\sigma[\Lambda, W]$.

The goal of this section is to characterize MEMDPs among POMDPs. To do so, we study the belief in PODMPs. When playing in a POMDP, we start with an initial belief about the current state[2]; that belief is updated according to the actions played and observations gathered.

▶ **Definition 18** (Belief in POMDP)**.** *Consider a POMDP $\Lambda = (Q, A, \delta, \Omega, O)$. A* belief *is a probability distribution $b \in \mathcal{D}(Q)$. Given $b \in \mathcal{D}(Q)$, $a \in A$, and $o \in \Omega$, we let $p(b,a,o) \in [0,1]$ denote the likelihood of $o$ given $(b,a)$:*

$$p(b,a,o) := \sum_{q\in O^{-1}(o)} \sum_{q'\in Q} b(q') \cdot \delta(q',a)(q)$$

[2] In POMDPs, we always have an initial belief about the current state. That is why, although MEMDPs in the "prior" semantics are special kinds POMDPs, it is not the case of MEMDPs in the "universal" semantics.

*We let* $\mathsf{Comp}(b,a) := \{o \in \Omega \mid p(b,a,o) > 0\} \neq \emptyset$ *denote the set of observations* compatible *with* $(b,a)$*. For all* $o \in \mathsf{Comp}(b,a)$*, we let* $\lambda[b,a,o] \in \mathcal{D}(Q)$ *denote the updated belief, defined by for* $q \in Q$*:*

$$\lambda[b,a,o](q) := \begin{cases} 0 & \text{if } O(q) \neq o \\ \frac{1}{p(b,a,o)} \cdot \sum_{q' \in Q} b(q') \cdot \delta(q',a)(q) & \text{otherwise} \end{cases}$$

When playing in an MEMDP, if we ever know for sure the current environment (i.e. the environment belief is Dirac), then this will never change. POMDPs induced by MEMDPs are thus said to be *Dirac-preserving*, i.e. for all $(q,a) \in Q \times A$ (with $q$ seen as a Dirac belief), we have: $\forall o \in \mathsf{Comp}(q,a) : |\mathsf{Supp}(\lambda[q,a,o])| = 1$.

As pointed out in [4], there is an alternative way to capture the behavior of POMDPs induced by MEMDPs via the (information theory) notion of entropy [15] of a belief. Formally, the entropy $H(b)$ of a belief $b \in \mathcal{D}(Q)$ is defined by $H(b) := -\sum_{q \in Q} b(q) \log(b(q)) \geq 0$. The entropy of a belief is a measure of the amount of information carried by that belief: the higher the entropy, the less carried information. A belief is Dirac if and only if its entropy is null; the maximal value of the entropy is $\log(|Q|)$, it is achieved by the uniform probability distribution.

In MEMDPs, the amount of information we have about the current environment never shrinks: the more distinguishing state-action pairs we visit, the better we know in which environment we are playing. As established in [4], this corresponds to the fact that POMDPs induced by MEMDPs have *non-increasing (expected) entropy* i.e. they are such that, for all $b \in \mathcal{D}(Q)$ and $a \in A$: $H(b) \geq \sum_{o \in \mathsf{Comp}(b,a)} p(b,a,o) \cdot H(\lambda[q,a,o])$.

Clearly, POMDPs with non-increasing entropy are Dirac-preserving. In fact, the converse is also true: all Dirac-preserving POMDPs have non-increasing entropy. (This is harder to show.)

▶ **Proposition 19.** *A POMDP has non-increasing entropy iff it is Dirac-preserving.*

The benefit of the above proposition is twofold. First, the notion of POMDPs with non-increasing entropy is natural, as this corresponds to POMDPs where the knowledge about the current state is monotonous (on average). However, it is not clear from the definition of POMDPs with non-increasing entropy how to effectively decide if a POMDP satisfies this property. It is now apparent that this can be done in polynomial time, since checking that a POMDP is Dirac-preserving can be done by enumerating all triplets of states, actions and observations. Second, given any Dirac-preserving POMDP with an observation-compatible parity objective[3], that is, a parity objective induced by a labelling function $f : Q \to \mathbb{N}$ such that $f(q_1) = f(q_2)$ whenever $O(q_1) = O(q_2)$, and given an initial belief, we can construct an exponentially larger MEMDP, together with a parity objective and an prior environment belief, such that the value in the POMDP coincides with the $\mathsf{pr}$-value in the MEMDP. This is formally stated below.

▶ **Theorem 20.** *Consider a Dirac-preserving POMDP* $\Lambda = (Q, A, \delta, \Omega, O)$*, an observation-compatible parity objective* $W \in \mathsf{Borel}(Q)$*, and an initial belief* $b \in \mathcal{D}(Q)$*. Then, we can compute in exponential time an MEMDP* $\Gamma = (Q', A, E, (\delta_e)_{e \in E})$*, a parity objective* $W' \in \mathsf{Borel}(Q')$*, and an environment belief* $b' \in \mathcal{D}(E)$ *such that there is a distinguished state* $q' \in Q'$ *for which* $\mathsf{val}(\Lambda, b, W) = \mathsf{val}^{\mathsf{pr}}_{q'}(\Gamma, b', W')$*.*

[3] Observation-compatible parity objectives, or visible objectives, are a common assumption in POMDPs. They define objectives that can be observed by the player: after playing and observing the sequence of observations traversed during the play, the player can determine whether it is winning or not; see for instance [5].

**Proof sktech.** Consider a Dirac-preserving POMDP $\Lambda = (Q, A, \delta, \Omega, O)$ and an initial belief $b \in \mathcal{D}(Q)$. Playing in a Dirac-preserving POMDP where the initial state is known amounts to playing in an MDP; similarly, playing in a MEMDP where the operating environment is known amounts to playing in a MDP. Thus, in the MEMDP $\Gamma$ that we build, we choose as set of environments the set of possible initial states, i.e. $E := \mathsf{Supp}(b)$, and we view the initial belief $b \in \mathcal{D}(Q)$ as the prior belief on the operating environment. As set of states, we consider the tuples of one current state in $Q$ per initial state in $\mathsf{Supp}(b) = E$ (this is where the exponential blow-up comes from); with all the states in a tuple mapped to the same observation in $\Omega$ by the function $O$. For all $e \in E = \mathsf{Supp}(Q)$, the definition of $\delta_e$ then relies on the Dirac-preserving assumption. Finally, an observation-compatible parity objective $W$ can then be translated into a parity objective on these tuples since the labeling function of $W$ only depends on the image of the states by the observation function $O$. ◀

We obtain that, up to an exponential blow-up, MEMDPs are exactly POMDPs with non-increasing entropy. Furthermore, note that the limit-sure [8] and gap [11] decision problems on arbitrary POMDPs with observation-compatible parity objectives are undecidable. Thus, Dirac-preserving POMDPs constitute a particularly well-behaved subclass of POMDPs.

## References


1 Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT Press, 2008.

2 Benjamin Bordais and Jean-François Raskin. Multi-environment mdps with prior and universal semantics. *CoRR*, abs/2602.10938, 2026. doi:10.48550/arXiv.2602.10938.

3 Eline M. Bovy, Caleb Probine, Marnix Suilen, Ufuk Topcu, and Nils Jansen. Code for the AB-HSVI algorithm and the experiments in the paper: "multi-environment pomdps: Discrete model uncertainty under partial observability" (neurips 2025) (version 1). https://doi.org/10.5281/zenodo.17425571, October 2025. Accessed on YYYY-MM-DD. doi:10.5281/ZENODO.17425571.

4 Krishnendu Chatterjee, Martin Chmelík, Deep Karkhanis, Petr Novotný, and Amélie Royer. Multiple-environment markov decision processes: Efficient analysis and applications. In J. Christopher Beck, Olivier Buffet, Jörg Hoffmann, Erez Karpas, and Shirin Sohrabi, editors, *Proceedings of the Thirtieth International Conference on Automated Planning and Scheduling, Nancy, France, October 26–30, 2020*, pages 48–56. AAAI Press, 2020. URL: https://ojs.aaai.org/index.php/ICAPS/article/view/6644.

5 Krishnendu Chatterjee, Laurent Doyen, and Thomas A. Henzinger. Qualitative analysis of partially-observable markov decision processes. In Petr Hlinený and Antonín Kucera, editors, *Mathematical Foundations of Computer Science 2010, 35th International Symposium, MFCS 2010, Brno, Czech Republic, August 23-27, 2010. Proceedings*, volume 6281 of *Lecture Notes in Computer Science*, pages 258–269. Springer, 2010. doi:10.1007/978-3-642-15155-2_24.

6 Krishnendu Chatterjee, Laurent Doyen, Jean-François Raskin, and Ocan Sankur. The value problem for multiple-environment mdps with parity objective. In Keren Censor-Hillel, Fabrizio Grandoni, Joël Ouaknine, and Gabriele Puppis, editors, *52nd International Colloquium on Automata, Languages, and Programming, ICALP 2025, Aarhus, Denmark, July 8–11, 2025*, volume 334 of *LIPIcs*, pages 150:1–150:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPIcs.ICALP.2025.150.

7 Nathanaël Fijalkow, Arka Ghosh, Roman Kniazev, Guillermo A. Pérez, and Pierre Vandenhove. Computing the reachability value of posterior-deterministic pomdps, 2026. doi:10.48550/arXiv.2602.07473.

8 Hugo Gimbert and Youssouf Oualhadj. Probabilistic automata on finite words: Decidable and undecidable problems. In Samson Abramsky, Cyril Gavoille, Claude Kirchner, Friedhelm Meyer auf der Heide, and Paul G. Spirakis, editors, *Automata, Languages and Programming, 37th International Colloquium, ICALP 2010, Bordeaux, France, July 6-10, 2010, Proceedings, Part II*, volume 6199 of *Lecture Notes in Computer Science*, pages 527–538. Springer, 2010. doi:10.1007/978-3-642-14162-1_44.

**9** Oded Goldreich. On promise problems: A survey. In Oded Goldreich, Arnold L. Rosenberg, and Alan L. Selman, editors, *Theoretical Computer Science, Essays in Memory of Shimon Even*, volume 3895 of *Lecture Notes in Computer Science*, pages 254–290. Springer, 2006. doi:10.1007/11685654_12.

**10** Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

**11** Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1-2):5–34, 2003. doi:10.1016/S0004-3702(02)00378-8.

**12** Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987. doi:10.1287/moor.12.3.441.

**13** Jean-François Raskin and Ocan Sankur. Multiple-environment markov decision processes. In Venkatesh Raman and S. P. Suresh, editors, *34th International Conference on Foundation of Software Technology and Theoretical Computer Science, FSTTCS 2014, New Delhi, India, December 15–17, 2014*, volume 29 of *LIPIcs*, pages 531–543. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2014. doi:10.4230/LIPIcs.FSTTCS.2014.531.

**14** Jean-François Raskin and Ocan Sankur. Multiple-environment markov decision processes. *CoRR*, abs/1405.4733, 2014. arXiv:1405.4733.

**15** Claude E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001. doi:10.1145/584091.584093.

**16** Maurice Sion. On general minimax theorems. *Pacific Jouronal of Mathematics*, 1958.

**17** Marnix Suilen, Marck van der Vegt, and Sebastian Junges. A PSPACE algorithm for almost-sure rabin objectives in multi-environment mdps. In Rupak Majumdar and Alexandra Silva, editors, *35th International Conference on Concurrency Theory, CONCUR 2024, Calgary, Canada, September 9–13, 2024*, volume 311 of *LIPIcs*, pages 40:1–40:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi:10.4230/LIPIcs.CONCUR.2024.40.

**18** John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev, 1947.