An Introduction to Multi-Environment Markov Decision Processes

Jean-François Raskin
Université libre de Bruxelles, Brussels, Belgium

Abstract

Markov Decision Processes (MDPs) are the standard model for decision-making under stochastic uncertainty. With full observability, reachability, safety and parity problems admit efficient algorithms and memoryless pure optimal strategies. Full observability is, however, often unrealistic. Partially Observable MDPs (POMDPs) replace it by an observation function, but the price is steep: most natural problems become undecidable. This motivates the study of structured subclasses or variants of POMDPs that preserve enough modelling power while restoring decidability. Multi-Environment MDPs (MEMDPs) are one such variant. An MEMDP is a finite family of MDPs with the same state and action spaces but distinct transition functions, and the active MDP, called the environment, is fixed throughout the run but hidden from the controller. The goal is to synthesize a single strategy that works well in every environment. Two semantics have been considered for MEMDPs: a *universal*, worst-case one, and a *prior*, Bayesian one based on a distribution over environments. We survey the main results obtained for both semantics since the introduction of the model in 2014. We cover reachability, parity and Rabin objectives, under the qualitative criteria (almost-sure, limit-sure) and the quantitative value-threshold problem, and we outline the key algorithmic ideas.

2012 ACM Subject Classification Theory of computation → Verification by model checking; Theory of computation → Logic and verification; Mathematics of computing → Markov processes

Keywords and phrases Multi-Environment Markov Decision Processes, partially observable MDPs, qualitative analysis, universal semantics, prior semantics, parity objectives

Digital Object Identifier 10.4230/LIPIcs.CONCUR.2026.4

Category Invited Talk

Funding The author of this work is funded by the FNRS and the ULB Fondation.

Acknowledgements I thank my co-authors Benjamin Bordais, Krishnendu Chatterjee, Laurent Doyen, and Ocan Sankur for many fruitful discussions on multi-environment MDPs, and for their technical contributions without which a large number of the results presented in this survey would not exist.

1 Introduction

Decision-making under uncertainty: MDPs and POMDPs. Markov Decision Processes (MDPs) are the central model for decision-making under stochastic uncertainty. At every step, a controller chooses an action and the system responds with a probabilistic transition to a successor state. When the state is fully observable, the standard problems (e.g. reachability, parity, mean-payoff) admit efficient algorithms and memoryless pure optimal strategies.

But full observability is often unrealistic. Partially Observable MDPs (POMDPs) replace the fully observable state by an observation of it: at each step, the controller selects an action based on the sequence of observations it has received so far, and the system then evolves as in an MDP, but the controller sees only the observation of the new state, not the state itself. In this richer setting, the controller may no longer be able to determine the current state with certainty. Instead, it maintains a *belief*, a probability distribution over states summarising everything the history of observations reveals about the hidden state. The belief is updated


© Jean-François Raskin;
licensed under Creative Commons License CC-BY 4.0

by Bayes' rule from the previous belief, the action played, and the observation received. In POMDPs, an optimal strategy can always be expressed as a function of the current belief alone. The belief space is, however, infinite and cannot be constructed effectively in general.

The expressive power of POMDPs comes at a cost. Most natural verification problems are *undecidable*: this is the case already for almost-sure and limit-sure reachability under infinite-memory strategies, for quantitative reachability, and for ω-regular objectives in general [7]. When decidability is recovered, complexities are typically high.

This has motivated the study of structured subclasses of POMDPs that preserve significant modelling power while restoring decidability. Multi-Environment MDPs are one such subclass.

Multi-Environment MDPs. A Multi-Environment Markov Decision Process (MEMDP), first introduced in [14], is a finite family of MDPs that share the same state space and the same action alphabet but differ in their transition functions. Initially, one environment is selected and then remains fixed for the rest of the run. The controller observes states and plays actions as in a plain MDP, but it never sees the index of the active environment. The goal is to synthesize a single strategy that performs well *no matter which environment was selected*.

Two semantics: universal and prior. Two natural semantics have been proposed for MEMDPs, depending on how the initial choice of environment is interpreted. The *universal semantics* was introduced in [14], the *prior semantics* in [5].

In the *universal semantics*, no distribution over environments is given. A strategy *wins* (resp. *guarantees value* $\geq v$) if it wins (resp. achieves value $\geq v$) in *every* environment. This is the worst-case viewpoint and the one adopted in most work on robust controller synthesis.

In the *prior semantics*, a probability distribution μ over environments is part of the input. The probability of winning, and the value of a strategy, are then averaged against μ. This is the Bayesian viewpoint: the controller starts with prior μ and refines a *belief* as observations accumulate. Theoretical properties of this semantics are studied systematically in [2].

Under the prior semantics, an MEMDP can be viewed as a POMDP whose states are of the form (s, i), where s is a state of the MEMDP, and i is the index of the operating environment. The index i is the only piece of hidden information, and it is sampled once at the beginning and then kept fixed. This immutability of the hidden component is precisely what makes MEMDPs with a prior significantly better behaved than general POMDPs: the controller can progressively infer which environment it is in, while the environment itself does not adapt in response to the controller's learning.

In contrast, under the universal semantics, the environment is selected in an adversarial manner, so an MEMDP under universal semantics is no longer a POMDP but rather a partially observable stochastic game. In this game, the adversary chooses the environment once at the start of the run, never changes it, and never discloses it; this initial hidden choice is the sole function of the adversary.

Applications. MEMDPs fit two complementary settings. The first is *controller synthesis against a finite set of adversaries.* Each adversary is a stochastic finite-memory strategy. Each strategy yields one environment. A winning MEMDP strategy wins against every adversary at once. The second is *robust control under model uncertainty.* The true transition function is one of finitely many candidates. The MEMDP collects them all. A winning strategy is robust to each of them.

[5] gives three concrete uses. In *recommender systems*, customer behaviour is an MDP. Clustering customers by attributes (age, gender, past choices) gives one environment per cluster. In *parametric MDPs*, some transition probabilities depend on unknown parameters with a discrete finite range. Each combination of parameter values is one environment. In *robot navigation under structural uncertainty*, the robot must reach a goal in a maze. The layout, or the action-failure probability, is unknown but drawn from a finite pool. Each candidate is one environment.

In all cases MEMDPs are expressive enough to model the uncertainty, yet structured enough to keep the analysis problems decidable and often tractable.

Structure of the paper. Section 2 fixes notation: MDPs and their strategies, MEMDPs, the two semantics, the objectives we consider (reachability, safety, parity, Rabin) and the decision problems (almost-sure, limit-sure, value-threshold, gap). Section 3 surveys results under the universal semantics and the main algorithmic ideas behind them (revealing edges, knowledge construction, distinguishing end-components). Section 4 turns to the prior semantics: we recall that the qualitative criteria coincide with the universal ones under full-support priors, present the truncated-belief construction of [2], obtain from it the presently strongest algorithm, in terms of worst-case complexity, for universal-value approximation, and discuss the connection with Dirac-preserving POMDPs. Section 5 reviews other related work, including robust MDPs.

2 Formal Definitions

This section fixes the notation for MDPs, MEMDPs, the two semantics, the objectives we consider, and the decision problems we study.

2.1 MDPs, strategies, objectives

For a finite or countable set X, $\mathcal{D}(X)$ denotes the set of probability distributions over X, and $\mathrm{supp}(\mu) \coloneqq \{x \in X : \mu(x) > 0\}$ is the support of $\mu \in \mathcal{D}(X)$. In this paper, we only consider distributions with *finite* support.

▶ **Definition 1** (MDP). *A (finite)* Markov Decision Process *is a tuple* $\mathcal{M} = (S, A, \delta, s_0)$ *where* S *is a finite set of states,* A *is a finite set of actions,* $s_0 \in S$ *is the initial state and* $\delta \colon S \times A \to \mathcal{D}(S)$ *is the probabilistic transition function. A triple* (s, a, s') *is an* edge *of the MDP if* $\delta(s, a)(s') > 0$.

A *randomized strategy* maps state histories to distributions over actions, $\sigma \colon S^+ \to \mathsf{Dist}(A)$, a *pure strategy* is a function $\sigma \colon S^+ \to A$, and a strategy is *finite-memory* if it can be realised by a (deterministic or stochastic) finite Mealy machine, see e.g. [7] for a formal definition. We write Σ for the set of all strategies. An initial state and a strategy induce a probability measure $\mathbb{P}^{\sigma}_{s_0}$ on infinite runs $s_0 s_1 s_2 \dots$ via a Markov chain and the classical cylinder construction.

▶ **Definition 2** (Markov chain). *A* Markov chain *is a tuple* $\mathcal{C} = (Q, P, q_0)$ *where* Q *is a finite or countable set of states,* $q_0 \in Q$ *is an initial state, and* $P \colon Q \to \mathcal{D}(Q)$ *is a stochastic transition function. A Markov chain induces, via the usual cylinder construction, a unique probability measure* $\mathbb{P}_{\mathcal{C}}$ *on the set* Q^{ω} *of infinite runs starting in* q_0*: for every finite prefix* $q_0 q_1 \dots q_n \in Q^+$*, the measure of the cylinder generated by that prefix equals* $\prod_{t=0}^{n-1} P(q_t)(q_{t+1})$*. The interested reader is referred to [1] for details on the cylinder construction.*

▶ **Definition 3** (Markov chain induced by a strategy). *Let $\mathcal{M} = (S, A, \delta, s_0)$ be an MDP and let $\sigma\colon S^+ \to \mathcal{D}(A)$ be a (randomized) strategy. The* Markov chain induced by σ from s_0 *is defined as $\mathcal{M}^\sigma_{s_0} \coloneqq (S^+, P^\sigma, s_0)$, where for every history $h \in S^+$ that ends in some state s and for all $s' \in S$, $P^\sigma(h)(h \cdot s') \coloneqq \sum_{a \in A} \sigma(h)(a) \cdot \delta(s, a)(s')$. The probability measure $\mathbb{P}^\sigma_{s_0}$, induced by the strategy σ, on infinite runs of $\mathcal{M}$ is obtained as the projection of $\mathbb{P}_{\mathcal{M}^\sigma_{s_0}}$ onto S^ω. Furthermore, this construction collapses to a finite Markov chain over S (resp. over $S \times \mathbb{M}$, where $\mathbb{M}$ is a finite memory set) whenever σ is memoryless (resp. a finite-memory strategy using memory $\mathbb{M}$).*

Objectives. An *objective* is a measurable set $\Omega \subseteq S^\omega$, i.e. a measurable set of infinite paths. Given an infinite run $\pi \in S^\omega$, we note $\mathsf{visit}(\pi)$ for the set of states visited by π, and $\mathsf{inf}(\pi)$ for the set of states visited infinitely often along π. We consider *reachability* $\mathsf{Reach}(T) \coloneqq \{\pi : \mathsf{visit}(\pi) \cap T \neq \emptyset\}$ for some target set $T \subseteq S$, *safety* $\mathsf{Safe}(T) \coloneqq \{\pi : \mathsf{visit}(\pi) \subseteq T\}$ for some set of safe states $T \subseteq S$, *parity* $\mathsf{Parity}(p)$ where $p\colon S \to \mathbb{N}$ is a priority function and runs are winning iff the *maximum* priority occurring infinitely often is even, i.e. $\mathsf{Parity}(p) = \{\pi : \max\{p(s) \mid s \in \mathsf{inf}(\pi)\}$ is even $\}$, and *Rabin* with pairs $((E_i, F_i))_{i \in I}$ where a run is winning iff, for some i, E_i is visited infinitely often while F_i is visited only finitely often, i.e. $\mathsf{Rabin}((E_i, F_i))_{i \in I} = \{\pi : \exists i \in I : \mathsf{inf}(\pi) \cap E_i \neq \emptyset \wedge \mathsf{inf}(\pi) \cap F_i = \emptyset\}$. Parity subsumes reachability, safety, and other classical objectives like Büchi, coBüchi, and is subsumed by Rabin. All the objectives defined here are measurable as they are all Borel objectives, for details see [18]. Given an MDP $\mathcal{M}$, a strategy σ, an initial state s_0, and an objective Ω, we denote by $\mathcal{M}^\sigma_{s_0}$, recall that the Markov chain induced by the strategy σ on $\mathcal{M}$ from its initial state, and so $\mathbb{P}_{\mathcal{M}^\sigma_{s_0}}(\Omega)$ denotes the probability of the set of infinite paths Ω when the strategy σ is played in $\mathcal{M}$.

2.2 MEMDP syntax

▶ **Definition 4** (MEMDP). *A* Multi-Environment MDP *is a tuple $\mathcal{E} = \big(S, A, (\delta_i)_{i \in \mathcal{I}}, s_0\big)$ where S, A, s_0 are as for plain MDP, $\mathcal{I}$ is a non-empty* finite *set of* environments*, and each $\delta_i\colon S \times A \to \mathcal{D}(S)$ is a transition function.*

For $i \in \mathcal{I}$, $\mathcal{M}_i \coloneqq (S, A, \delta_i, s_0)$ is the MDP of environment i. The controller observes states but not the environment index, so any two environments are observationally equivalent on state information alone. We write $\mathbb{P}^\sigma_i(\Omega)$ for the probability of Ω under strategy σ in environment i.

2.3 Universal semantics

In the universal semantics, no distribution over $\mathcal{I}$ is given: the environment is selected adversarially.

▶ **Definition 5** (Universal semantics – winning conditions). *A strategy σ is* universally almost-sure winning *for Ω if $\mathbb{P}^\sigma_i(\Omega) = 1$ for every $i \in \mathcal{I}$. The MEMDP $\mathcal{E}$ is* universally limit-sure winning *for Ω if $\sup_\sigma \min_{i \in \mathcal{I}} \mathbb{P}^\sigma_i(\Omega) = 1$: for every $\varepsilon > 0$ there exists σ achieving Ω with probability at least $1 - \varepsilon$ in every environment. The* universal value *is $\mathsf{val}^U(\Omega) \coloneqq \sup_\sigma \min_{i \in \mathcal{I}} \mathbb{P}^\sigma_i(\Omega)$. A strategy σ solves the* universal λ-threshold problem *if $\mathbb{P}^\sigma_i(\Omega) \geq \lambda$ for every $i \in \mathcal{I}$.*

Three guiding examples. Figure 1 shows three MEMDPs from [14] that illustrate key properties under the universal semantics. Each MEMDP has two environments M_1 and M_2. Dashed edges exist only in M_1, dotted edges only in M_2, and stochastic branchings are uniform unless specified otherwise. We use simple reachability objectives, with target T, and absorbing sinks are made explicit by self-loops.

The MEMDP of Figure 1a shows that *randomization is, in general, necessary* even for reachability objectives. In M_1, action a from s reaches the target u deterministically, while b reaches the non-target t, and in M_2 the two actions are swapped. Any pure strategy wins in exactly one environment, so the best pure strategy has *universal* value 0. But, playing a and b each with probability $\frac{1}{2}$ reaches T with probability $\frac{1}{2}$ in both environments, which is the universal value of this MEMDP.

The MEMDP of Figure 1b shows that *infinite memory may be required*. Once u is reached, the controller must commit: the winning action is b in M_1 and a in M_2. To pick the right action, it can repeatedly play a at s and record the empirical frequencies of the self-loop on s and of visits to t. State u is reached with probability 1 under every strategy, so the controller indeed has to commit there. The inference of the operating environment can be modeled as a *binary hypothesis test*. Assume that from s under action a the transition probabilities to u, s and t are $(1-p_1-p_2, p_1, p_2)$ in M_1, $(1-p_1'-p_2', p_1', p_2')$ in M_2. After n visits to s under a, let $\widehat{p}_1$ be the empirical frequency of the self-loop on s. The controller can apply the following rule. Assume "M_1" iff $\widehat{p}_1$ is closer to p_1 than to p_1'. By Hoeffding's inequality[1], the probability of misclassification decay exponentially in n. Still, to drive the misclassification probability below an arbitrary $\varepsilon > 0$, the number of samples needed diverges as $\varepsilon \to 0$, so any optimal strategy uses *unbounded memory*.

The MEMDP of Figure 1c shows that the value 1 *need not be attained*, even when it can be approached arbitrarily closely, showing that optimal strategies need not exist. Looping on a between s and t reveals statistical information about the operating environment without leaving $\{s, t\}$. From s, action b or c leads to the target v or to the non-target u, but which action is correct depends on the environment: c in M_1, b in M_2. For every $\varepsilon > 0$, a finite-memory strategy samples a enough times (according to Hoeffding's inequality) and then commits to its best estimate, reaching T with probability at least $1-\varepsilon$ in both environments. The value 1 itself, however, is not attained: any guess fails with positive probability. So according to the winning conditions introduced in Def. 5, the MEMDP is limit-sure winning but not almost-sure winning.

Together, these three examples already exhibit the main subtleties of the universal semantics: the role of randomization, the gap between finite and infinite memory, and the gap between almost-sure and limit-sure winning.

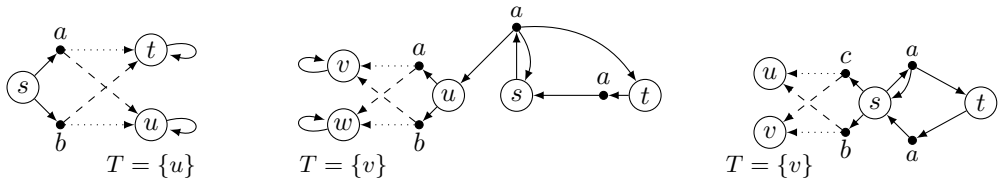
2.4 Prior semantics

In the prior semantics, besides the MEMDP $\mathcal{E}$, the input also includes a distribution $\mu \in \mathcal{D}(\mathcal{I})$ with rational coefficients. The environment is modeled as a random variable distributed according to μ, which encodes prior information about which environment is currently active.

[1] Hoeffding's inequality provides a bound on the probability that the sum of random variables X_i deviates from its expected value. Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [0,1]$, and let $\epsilon > 0$ be a deviation, then:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right) \tag{1}$$

This inequality helps us determine the number n of samples needed to ensure the empirical average is within ϵ of the expected average with high probability.



(a) Randomization is necessary. **(b)** Infinite memory is necessary. **(c)** Limit-sure ≠ almost-sure.

Figure 1 Three MEMDPs from [14]. Dashed edges exist only in M_1, dotted edges only in M_2; stochastic branchings are uniform unless specified otherwise. (a) Randomization is necessary: the universal value is $\frac{1}{2}$, attained by uniform randomization between a and b. (b) Sampling a at s lets the controller infer the active environment with *high probability*; an optimal strategy at uv needs to remember the unbounded history up to there, it require thus infinite memory. (c) Target $T = \{v\}$ is reached with probability $1 - \varepsilon$ for every $\varepsilon > 0$, but not almost surely: limit-sure ≠ almost-sure.

▶ **Definition 6** (Prior semantics). *Given a prior μ, define $\mathbb{P}^\sigma_\mu(\Omega) \coloneqq \sum_{i\in\mathcal{I}} \mu(i) \cdot \mathbb{P}^\sigma_i(\Omega)$. A strategy σ is* prior almost-sure winning *iff $\mathbb{P}^\sigma_\mu(\Omega) = 1$; the MEMDP is* prior limit-sure winning *iff $\sup_\sigma \mathbb{P}^\sigma_\mu(\Omega) = 1$. The* prior value *is $\mathsf{val}^P_\mu(\Omega) \coloneqq \sup_\sigma \mathbb{P}^\sigma_\mu(\Omega)$. A strategy solves the* prior λ-threshold problem *if $\mathbb{P}^\sigma_\mu(\Omega) \geq \lambda$.*[2]

▶ **Example 7.** Consider the three MEMDPs of Figure 1 under the uniform prior $\mu(M_1) = \mu(M_2) = \frac{1}{2}$.

In Figure 1a, the pure strategy "play a at s" has universal value 0 but prior value $\mu(M_1) \cdot 1 + \mu(M_2) \cdot 0 = \frac{1}{2}$, matching the randomized strategy "play a or b uniformly". Randomization is therefore not required under the prior semantics, consistent with Proposition 19 below: MEMDPs with a prior are POMDPs, for which pure strategies with enough memory are optimal for all the classes of objectives that we consider in this paper.

In Figure 1b, an optimal strategy still uses the different distributions of a at s to infer the active environment, but the inference is now expressed as a *Bayesian belief update*[3] on $\mathcal{I}$: after a small number of observations the belief $b_t \in \mathcal{D}(\mathcal{I})$ concentrates on the more likely environment, and the controller commits to the matching action at u. This is the intuition behind the truncated-belief construction of Section 4.

In Figure 1c, the two semantics agree on the qualitative side: the MEMDP is limit-sure but not almost-sure winning in both. This is the general phenomenon – under any prior of full support, the two semantics agree on the qualitative criteria [2].

2.5 Decision problems

For a class of objectives $\mathcal{C} \in \{\mathsf{Reach}, \mathsf{Safe}, \mathsf{Parity}, \mathsf{Rabin}\}$ and a semantics $\tau \in \{U, P\}$, we consider the following problems. Fix an MEMDP $\mathcal{E}$, and a prior μ in the prior case.

▶ **Problem 8** (Qualitative). Almost-sure *(resp.* limit-sure*)* winning*: given $\Omega \in \mathcal{C}$, decide whether some strategy is almost-sure (resp. limit-sure) winning in semantics τ.*

▶ **Problem 9** (Value-threshold / gap). *Given $\Omega \in \mathcal{C}$, a rational $v \in [0, 1]$ and $\varepsilon > 0$:*

- *output YES if $\mathsf{val}^\tau(\Omega) \geq v$;*
- *output NO if $\mathsf{val}^\tau(\Omega) \leq v - \varepsilon$;*
- *either answer is acceptable if $\mathsf{val}^\tau(\Omega) \in (v - \varepsilon, v)$.*

[2] We omit μ from the notation when it is clear from context. Moreover, all concepts here are implicitly defined with respect to the initial state s_0; if a different state is considered, it will be indicated explicitly in the notation.

[3] We recall the notion of belief and its update later in the paper.

The gap problem is the natural approximate version of the value problem, the exact decidability of which is, to the best of our knowledge, open under both semantics, even for reachability.

3 Results under the Universal Semantics

We summarize the main decidability and complexity results for the universal semantics and sketch the core algorithmic methods behind them. The seminal work [14] introduced the model and proved decidability for the case of two environments. The paper [8] extends these results to n environments, in particular providing algorithms for limit-sure parity and the gap problem, while [16] settles the almost-sure Rabin case, and [17] presents corresponding results for reachability. We illustrate the central notions of revealing edges and distinguishing edge components using the examples in Figure 2 and Figure 3.

3.1 Almost-sure winning

▶ **Theorem 10** ([14, 8, 16]). *The almost-sure reachability, safety, parity and Rabin problems for MEMDPs under the universal semantics are PSPACE-complete. Finite-memory pure strategies suffice, with memory at most exponential in $|\mathcal{I}|$ and polynomial in $|S|$. For fixed $|\mathcal{I}|$, the problems are solvable in polynomial time.*

We sketch the algorithms that establish the upper-bounds.

Revealing edges. An edge (s, a, s') is *revealing* if the predicate $\delta_i(s, a)(s') > 0$ depends on the environment i. Whenever a revealing edge is traversed, we learn that the active environment satisfies that predicate. The current *knowledge* $K \subseteq \mathcal{I}$ is the set of environments still compatible with the observations made so far, and we note that crossing a revealing edge can only shrink K. Under the universal semantics, almost-sure winning amounts to gathering this information, and since $\mathcal{I}$ is finite, only finitely many distinct K's arise.

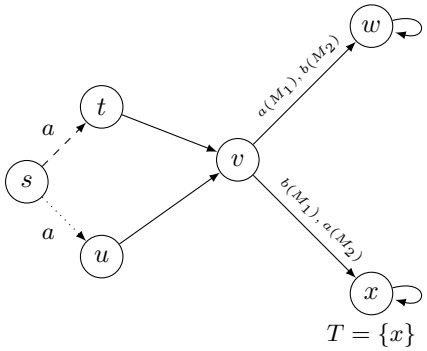


Figure 2 Illustration of revealing edges. Dashed edges exist only in M_1, dotted edges only in M_2; w is absorbing non-target and x is the absorbing target. Starting state is s. Every edge leaving s is revealing: the transition (s, a, t) (resp. (s, a, u)) reveals that the active environment is M_1 (resp. M_2). In M_1, in v, playing a leads to w, and playing b leads to the target x. Conversely, in M_2, in v, playing a leads to the target x, and playing b leads to w.

▶ **Example 11.** Consider the MEMDP of Figure 2, with target $T = \{x\}$, two environments M_1 (dashed) and M_2 (dotted), an absorbing non-target w and an absorbing target x. From s, action a leads to t in M_1 and to u in M_2: a single observation of the successor of s under

a identifies the active environment. If t is reached, the controller knows it plays in M_1 and from v plays b; if u is reached, it knows it plays in M_2 and from v plays a. One revealing transition collapses the knowledge $\{M_1, M_2\}$ to either $\{M_1\}$ or $\{M_2\}$, and each residual case is a plain MDP reachability problem. This is the phenomenon exploited by the knowledge construction below: almost-sure winning on the original MEMDP reduces, after traversing revealing transitions, to almost-sure winning on strictly smaller environment sets.

The knowledge construction. Build an MDP $\mathcal{E}^K$ whose states are pairs (s, K) with $K \subseteq \mathcal{I}$ the current knowledge. On action a and observed successor s', K updates to $K' = \{i \in K : \delta_i(s,a)(s') > 0\}$. A strategy is universally almost-sure winning in $\mathcal{E}$ iff its knowledge-lifted version is almost-sure winning in $\mathcal{E}^K$ viewed as a standard MDP. $\mathcal{E}^K$ has at most $|S| \cdot 2^{|\mathcal{I}|}$ states, which gives an exponential-time algorithm, and since K shrinks monotonically along any run, a recursive procedure on $|K|$ runs in polynomial space. For fixed $|\mathcal{I}|$, the construction is polynomial-time.

Lower bounds. [17] establishes a PSPACE lower bound for universal almost-sure reachability by reducing TQBF to it. Every component $\mathcal{M}_i$ in the reduction is acyclic, so every run reaches in finitely many steps a target or non-target sink. The same construction therefore yields a PSPACE lower bound for the almost-sure variant of safety, parity and Rabin: on acyclic MEMDPs these objectives all reduce to reachability on the absorbing tail.

3.2 Limit-sure winning

Limit-sure winning asks whether the value equals 1, without requiring the supremum to be attained. It is strictly weaker than almost-sure winning, as shown by Figure 1(c).

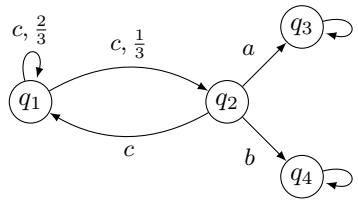
▶ **Theorem 12** ([14, 8]). *The limit-sure reachability, safety and parity problems for MEMDPs under the universal semantics are PSPACE-complete. Finite-memory pure strategies witness any ε-approximation. For fixed $|\mathcal{I}|$, the problems are in polynomial time.*

The algorithm reuses revealing edges and the recursion on knowledge, but also needs an adapted notion of end-component.

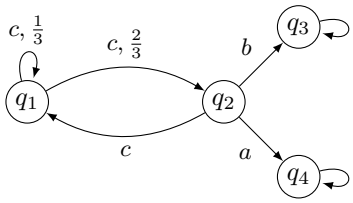
Distinguishing end-components. End-components (ECs) are central to MDP analysis [9, 1]: an EC is a strongly connected sub-MDP, and under any strategy a run almost surely settles in some EC, with each of its states and edges visited infinitely often when actions are played uniformly. In an MEMDP, transition supports can differ across environments, so the notion is adapted. A *common end-component* (CEC) for a set $K \subseteq \mathcal{I}$ is a pair (S', A') that induces an EC in every $i \in K$; this notion was introduced for two environments in [14] and extended to n in [8]. A CEC is *distinguishing* if it contains an edge (q_1, a, q_2) with $\delta_{i_1}(q_1, a)(q_2) \neq \delta_{i_2}(q_1, a)(q_2)$ for some $i_1, i_2 \in K$.

Staying long enough inside a distinguishing CEC lets the controller estimate, by the law of large numbers (based on Hoeffding's inequality), the transition probabilities of distinguishing edges with arbitrary confidence and error. The controller can therefore identify, with probability arbitrarily close to 1, the set of environments consistent with what it observes. We illustrate distinguishing CECs and a more subtle variant on two examples from [8].

▶ **Example 13** (A distinguishing CEC (Figure 3)). Two environments e_1, e_2; states $\{q_1, q_2, q_3, q_4\}$; target $T = \{q_3\}$; q_3 and q_4 absorbing. For $K = \{e_1, e_2\}$, the pair $D = (\{q_1, q_2\}, \{c\})$ is a CEC: from q_1, action c has support $\{q_1, q_2\}$ in both environments, and from q_2 action c returns to q_1.



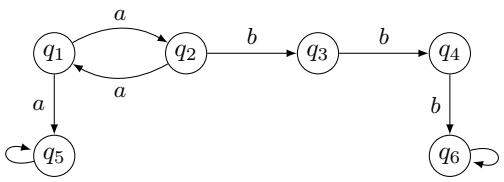
(a) In environment e_1.



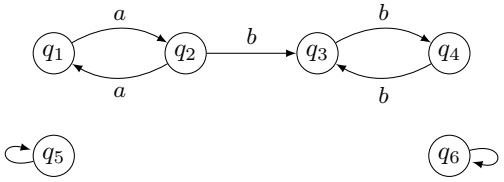
(b) In environment e_2.

Figure 3 A distinguishing common end-component. The set $D = \{q_1, q_2\}$ is an end-component in both environments, but action c at q_1 has different probability distributions in e_1 and e_2, so repeated use of c lets the controller infer the operating environment with arbitrarily high confidence.

Note that in this CEC, no single transition is revealing (the supports agree), so almost-sure winning is impossible: committing to a or b at q_2 fails in one environment, and staying inside the CEC D never reaches q_3. However, the distribution of c at q_1 differs: $\delta_{e_1}(q_1, c) = (\frac{2}{3}, \frac{1}{3})$ and $\delta_{e_2}(q_1, c) = (\frac{1}{3}, \frac{2}{3})$ on (q_1, q_2). Hence D is a distinguishing CEC: by playing c long enough and comparing empirical frequencies, the controller guesses the environment with error probability at most ε, then commits to a at q_2 in the "e_1-branch" and to b in the "e_2-branch". This shows q_1 is limit-sure but not almost-sure winning.



(a) In environment e_1: action a at q_1 may escape to q_5.



(b) In environment e_2: $\{q_1, q_2\}$ is an EC.

Figure 4 An end-component that is present only in some environments. Priority 0: $\{q_3, q_4, q_5\}$; priority 1 elsewhere. The set $\{q_1, q_2\}$ is an end-component in e_2 but *not* in e_1, where action a at q_1 can escape into the priority-0 trap q_5.

▶ **Example 14** (An end-component that is "escaped" in some environments (Figure 4).)**.** Distinguishing CECs are not the only structure enabling limit-sure winning. Consider the MEMDP of Figure 4, with two environments e_1, e_2, states $\{q_1, q_2, q_3, q_4, q_5, q_6\}$ and a parity objective of priority 0 on $\{q_3, q_4, q_5\}$ and priority 1 elsewhere. States q_4, q_5, q_6 are absorbing; q_5 is a priority-0 absorbing "good trap" and q_6 is a priority-1 absorbing "bad trap".

A key difference with Figure 3 is that the set $D = \{q_1, q_2\}$ is *not* a common end-component: in e_2 it is closed under action a, but in e_1 action a at q_1 has a positive probability to escape to q_5. The transition (q_1, a, q_2) itself is *not* revealing, as it is present in both environments; yet, after crossing it many times, the controller can still argue statistically that it is very likely in e_2 as in e_1, every visit to q_1 carries a fixed positive probability of exiting to q_5 when playing a, so not reaching q_5 after many rounds becomes arbitrarily unlikely in e_1. The limit-sure strategy at q_1 is therefore: *play a at q_1 and q_2 for a long time*, and one of two things happens:

- Either q_5 is eventually reached, which happens almost surely in e_1 and is winning there (priority 0 forever);
- or q_5 is not reached after sufficiently many rounds, in which case the controller concludes (with confidence $1 - \varepsilon$) that the environment is e_2, and switches to playing b at q_2 to reach the priority-0 state q_3 (and then q_4), satisfying the parity objective in e_2 as well.

Hence q_1 is limit-sure winning but not almost-sure winning. The structure exploited here is different from that of Figure 3: $D = \{q_1, q_2\}$ is an end-component in one environment only, and has "exiting" edges that are winning in the other environment. The limit-sure analysis of [8] therefore has to recognise, in addition to distinguishing common end-components, end-components that are broken by winning exits in some environments.

Lower bound. The PSPACE lower bound for almost-sure winning holds already for acyclic MEMDPs, on which limit-sure and almost-sure coincide, as a consequence the same lower bound applies to limit-sure winning.

3.3 Value approximation and the gap problem

The exact value problem is open under the universal semantics, for parity as well as for safety and reachability. The following approximation result is, however, available.

▶ **Theorem 15** ([8]). *For every MEMDP $\mathcal{E}$, parity objective Ω, threshold λ and precision $\varepsilon > 0$, the gap problem is decidable; i.e., one can decide whether $\mathsf{val}^U(\Omega) \geq \lambda$ or $\mathsf{val}^U(\Omega) \leq \lambda - \varepsilon$. Moreover, a strategy σ_ε enforcing universal value at least λ has size computable and bounded by a function of $|S|$, $|\mathcal{I}|$ and ε.*

The proof first shows that finite memory suffices to approximate the value, then reduces the gap problem to checking a formula in the theory of the reals whose size is bounded by the memory bound. A sharper complexity bound, going through the prior semantics, is summarised in Section 4.

▶ Remark 16. Unlike the almost-sure and limit-sure cases, randomization may be strictly necessary to achieve values $c < 1$ under the universal semantics [14]. For instance, in the example of Figure 1(a), randomization is required to attain value $\frac{1}{2}$.

4 Results under the Prior Semantics

The prior semantics is Bayesian in nature and makes MEMDPs compatible with the standard belief-state framework for POMDPs. In what follows, we successively revisit its agreement with the universal semantics for qualitative objectives, the truncated-belief construction of [2] for computing an approximation of the prior value, the connection between universal and prior values that allows us to refine approximations of the universal value, and a new class of well-behaved POMDP that we call Dirac-preserving POMDPs.

4.1 Agreement with the universal semantics for qualitative criteria

▶ **Proposition 17** ([2]). *Let $\mathcal{E}$ be an MEMDP, μ a prior of full support on $\mathcal{I}$, and Ω a parity objective. For every strategy σ, $\forall i \in \mathcal{I}$, $\mathbb{P}_i^\sigma(\Omega) = 1 \iff \mathbb{P}_\mu^\sigma(\Omega) = 1$. Similarly, $\sup_\sigma \mathbb{P}_\mu^\sigma(\Omega) = 1$ iff $\mathcal{E}$ is universally limit-sure winning.*

The intuition is immediate: under a full-support prior, probability 1 (resp. approaching 1) on average implies probability 1 (resp. approaching 1) in every environment, and vice versa.

4.2 MEMDPs with prior as POMDPs

We briefly review the POMDP framework and note that an MEMDP equipped with a prior can be seen as a POMDP of a specific form. We will return to this point later in the discussion of Dirac-preserving POMDPs.

▶ **Definition 18** (POMDP). *A (finite)* Partially Observable Markov Decision Process *is a tuple* $\mathcal{P} = \big(S, A, \delta, \mathcal{O}, \mathrm{obs}, \iota_0\big)$ *where* S *and* A *are finite sets of states and actions,* $\delta\colon S \times A \to \mathcal{D}(S)$ *is the probabilistic transition function,* $\mathcal{O}$ *is a finite set of observations,* $\mathrm{obs}\colon S \to \mathcal{O}$ *is an observation function, and* $\iota_0 \in \mathcal{D}(S)$ *is an initial distribution. Strategies for POMDPs depend only on the sequence of observations already produced, i.e., they are functions* $\sigma\colon \mathcal{O}^+ \to \mathcal{D}(A)$. *A POMDP together with a strategy and the initial distribution* ι_0 *defines a probability measure on* S^ω *via the classical cylinder construction and the corresponding Markov chain.*

▶ **Proposition 19** (MEMDPs with prior as POMDPs). *Let* $\mathcal{E} = (S, A, (\delta_i)_{i\in\mathcal{I}}, s_0)$ *be an MEMDP and* $\mu \in \mathcal{D}(\mathcal{I})$ *a prior. Define the POMDP* $\mathcal{P}_{\mathcal{E},\mu} := (\widehat{S}, A, \widehat{\delta}, \mathcal{O}, \mathrm{obs}, \widehat{\iota_0})$ *by*

- $\widehat{S} := S \times \mathcal{I}$ *(the original state space is duplicated, once per environment);*
- $\widehat{\delta}\big((s,i),a\big)\big((s',j)\big) := \delta_i(s,a)(s')$ *if* $j = i$, *and* 0 *otherwise (the environment index is static along any run);*
- $\mathcal{O} := S$ *and* $\mathrm{obs}(s,i) := s$ *(the controller observes the original state but not the environment);*
- $\widehat{\iota_0}(s,i) := \mu(i)$ *if* $s = s_0$, *and* 0 *otherwise (the only source of uncertainty is the initial choice of environment, distributed according to* μ, *but hidden to the controller).*

Then, for every objective Ω *on* S^ω *and every strategy* σ *of* $\mathcal{E}$ *(which, since states are fully observed, is already a POMDP strategy for* $\mathcal{P}_{\mathcal{E},\mu}$*),* $\mathbb{P}^\sigma_{\mathcal{P}_{\mathcal{E},\mu}}(\Omega) = \mathbb{P}^\sigma_\mu(\Omega)$. *In particular,* $\mathsf{val}^P(\mathcal{E},\mu) = \mathsf{val}(\mathcal{P}_{\mathcal{E},\mu})$.

An MEMDP with a prior is therefore a POMDP in which the hidden coordinate is *frozen* after the initial draw: only the initial environment is uncertain, and observations refine the belief about it but it never changes along the run.

Before turning to the MEMDP-specific algorithm, we recall the notion of *belief* for a general POMDP $\mathcal{P} = (S, A, \delta, \mathcal{O}, \mathrm{obs}, \iota_0)$. A belief is a probability distribution $b \in \mathcal{D}(S)$ over the hidden state. The initial belief is ι_0; on action a and observation $o \in \mathcal{O}$, b is updated by Bayes' rule to b', where $b'(s')$ is proportional to $\sum_s b(s) \cdot \delta(s,a)(s')$ if $\mathrm{obs}(s') = o$ and 0 otherwise. The reachable beliefs, with their induced transitions, form the (in general infinite) *belief MDP* of the POMDP; the value of the POMDP coincides with the value of its belief MDP. This construction is classical; see, e.g., [1]. Since the belief MDP is infinite in general, algorithms need either to rely on structural properties of the belief space or a finite abstraction of it. In the case of MEMDPs with prior, both apply: beliefs live in $\mathcal{D}(\mathcal{I})$ (which is potentially much simpler than $\mathcal{D}(S)$), their support only shrinks over time, and a controlled finite abstraction yields a ε-approximation of the value.

4.3 Truncated belief construction

For MEMDPs with a prior, the belief update specialises as follows. Given a belief $b \in \mathcal{D}(\mathcal{I})$, state s, action a and observed successor s':

$$\mathsf{Update}(b,s,a,s')(i) \;:=\; \frac{b(i)\,\delta_i(s,a)(s')}{\sum_{j\in\mathcal{I}} b(j)\,\delta_j(s,a)(s')} \qquad (i \in \mathcal{I}),$$

whenever the denominator is nonzero (i.e., s' is reachable from s via a under b). Starting from $b_0 = \mu$ and iterating Update along a run yields the trajectory $(b_t)_{t\geq 0}$ in $\mathcal{D}(\mathcal{I})$. As noted above, because the active environment is static, the set of reachable beliefs is infinite but highly structured.

The main result of [2, Sect. 3.2] is that the prior value can be approximated to any precision by a finite belief MDP. We summarise the four ingredients of the construction; each one points to a precise statement in [2].

(1) Close beliefs yield close values. Lemma 4 of [2] shows that the prior value is 1-Lipschitz in the belief: for every $b, b' \in \mathcal{D}(\mathcal{I})$ and state s, $\left|\mathsf{val}^P_{s,b}(\Omega) - \mathsf{val}^P_{s,b'}(\Omega)\right| \leq \frac{1}{2}\|b - b'\|_1$. Rounding the belief to a nearby representative therefore costs at most the rounding error on the value.

(2) Non-distinguishing transitions preserve the belief. Call (s, a, s') *non-distinguishing* for b if $\delta_i(s, a)(s')$ agrees across all i in the support of b. Then $\mathsf{Update}(b, s, a, s') = b$: new beliefs are reached only via *distinguishing* transitions, and any region of the MEMDP without such a transition contributes a single belief.

(3) Suppressing a small environment after enough distinguishing observations. Theorem 5 of [2] formalises the following idea: as distinguishing transitions accumulate, the belief concentrates, and with high probability some component $b_t(i)$ drops below an arbitrarily small threshold $\eta > 0$. The probability of a run along which every component stays above η decreases exponentially in the number of distinguishing transitions. Whenever a component falls below η, the belief is *pruned*: the corresponding environment is dropped and the remaining components are renormalised, at a cost bounded by η on the value (by step (1)). In the uncommon cases where every component remains above η, we may uniformly at random suppress a single environment. This can be done because this happens with an arbitrarily small probability whenever we have seen enough distinguishing transitions, see Theorem 5 of [2] for the formal details.

(4) A finite belief MDP that approximates the MEMDP. Combining (1)–(3), every reachable belief is, up to a controlled loss, equivalent to a rounding to a grid of resolution η, possibly composed with a pruning of components below η. The resulting beliefs form a finite subset of $\mathcal{D}(\mathcal{I})$ of size polynomial in $1/\eta$ for fixed $|\mathcal{I}|$. Pairs $(s, \widehat{b})$ form the states of a finite *belief MDP* whose value approximates $\mathsf{val}^P_\mu(\Omega)$ within ε for η polynomial in ε and $|\mathcal{I}|$. Optimal positional strategies on the finite belief MDP lift to ε-optimal finite-memory strategies of $\mathcal{E}$ whose memory size is polynomial in $1/\varepsilon$ for fixed $|\mathcal{I}|$; see [2, Sect. 3.2] for the precise statements.

It is important to note that no analogue of Theorem 5 holds for general POMDPs: the ε-approximation of the value of non-discounted infinite-horizon objectives is not computable [6]. The static nature of the hidden environment is precisely what makes steps (2) and (3) sound, and is the structural reason why the truncated-belief approach yields computable approximations for MEMDPs with prior.

▶ **Example 20.** We illustrate the belief update on the running example of [2], a card game modelled as the MEMDP of Figure 5. The deck has two card types, mixed with parameters $\alpha_1, \alpha_2 = 1 - \alpha_1$. At each turn the player can DRAW (continuing with probability $1 - \alpha_0$, forced to guess with probability α_0) or GUESS the duplicated card. Guessing i from observation C_j reaches W with probability β_i if $j = 1$ and γ_i if $j = 2$. The objective is reachability of W. The two environments are dual: in E_1, $\alpha_1 = \frac{2}{3}$, $\beta_1 = 1, \beta_2 = 0, \gamma_1 = 0, \gamma_2 = 1$; in E_2 the values of α_1 and α_2 are swapped, as are those of the guessing parameters.

Limit-sure, not almost-sure. Set $\alpha_0 = 0$. No strategy wins with probability exactly 1: after any finite number of draws the empirical frequencies may mislead the guess with small but nonzero probability. By the law of large numbers, however, for every $\varepsilon > 0$ enough draws push the probability of a misleading sample below ε, and then guessing the observed majority reaches W with probability at least $1 - \varepsilon$ in both environments. The MEMDP is limit-sure but not almost-sure winning. Since the two semantics agree qualitatively under full-support priors, the same holds in the prior semantics for any μ with $\mu(E_1), \mu(E_2) > 0$.

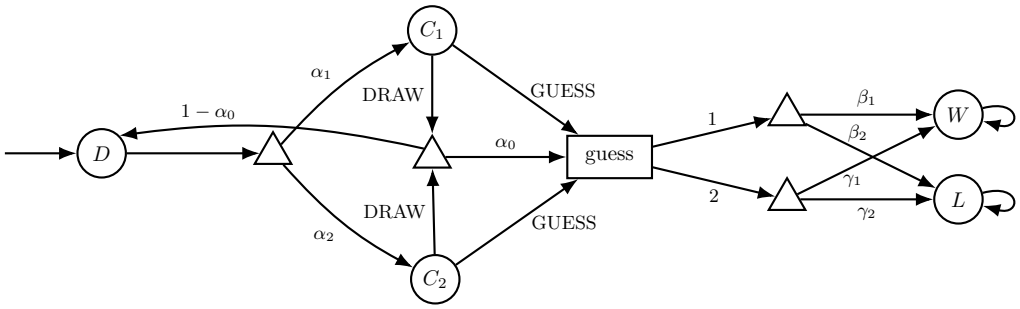


Figure 5 The duplicated-deck MEMDP from [2], inspired by [8]. A deck contains two card types with mixing parameters α_1, α_2 ($\alpha_1 + \alpha_2 = 1$). At each turn the player may request a DRAW (continuing with probability $1 - \alpha_0$, forced to guess with probability α_0) or directly GUESS which card is duplicated. The two environments E_1 and E_2 correspond to card 1 or card 2 being duplicated, obtained by swapping the numerical values of α_1, α_2 and of the guessing parameters β_i, γ_i.

Belief update along a run. Take $\mu = (\frac{1}{4}, \frac{3}{4})$, biased towards E_2. If the first draw yields card 2 (reaching observation C_2), the Bayes' update rule gives

$$\mathsf{Update}(\mu, D, \text{DRAW}, C_2)(E_2) = \frac{\frac{3}{4} \cdot \frac{2}{3}}{\frac{1}{4} \cdot \frac{1}{3} + \frac{3}{4} \cdot \frac{2}{3}} = \frac{6}{7}, \qquad \mathsf{Update}(\mu, D, \text{DRAW}, C_2)(E_1) = \frac{1}{7}.$$

If the first draw yields card 1, the posterior is $(\frac{2}{5}, \frac{3}{5})$: the evidence against E_2 is weaker than the evidence for it, reflecting the asymmetric prior. Iterating these updates yields a trajectory $(b_t)_{t \geq 0}$ that concentrates on a single environment, provided distinguishing observations occur often enough (here, every DRAW is distinguishing).

The truncated-belief construction of Section 4.3 turns this concentration into a finite abstraction: merge beliefs within distance η, prune components below η when the support has size ≥ 2, at a cost on the value controlled by Lemma 4 of [2].

4.4 Relationship between the two semantics

The universal and prior values are tightly linked, and this link is the basis of the currently best approximation algorithm for the universal value.

▶ **Theorem 21** (adapted from Theorem 7 of [2]). *For every MEMDP $\mathcal{E}$ and parity objective Ω, $\mathsf{val}^U(\Omega) = \inf_{\mu \in \mathcal{D}(\mathcal{I})} \mathsf{val}^P_\mu(\Omega)$.*

The inequality $\mathsf{val}^U \leq \mathsf{val}^P_\mu$ for every μ is immediate: a strategy of universal value $\geq v$ has prior value $\geq v$ for any prior. The reverse inequality goes through *mixed strategies*[4], i.e., distributions over pure strategies. To establish the theorem, we need to combine two ingredients: (*i*) a generalisation of von Neumann's min-max theorem (applicable because $\mathcal{I}$ is finite), the supremum-over-mixed-strategies of the universal value equals the infimum-over-priors of the supremum-over-mixed-strategies of the prior value; and (*ii*) mixed strategies are not more powerful than randomized strategies on the universal value.

Theorem 21 reduces approximation of the universal value to approximation of the prior value over a sufficient set of priors. By Lemma 4 of [2], the prior value is 1-Lipschitz in the belief, so the infimum over $\mathcal{D}(\mathcal{I})$ is approximated by the minimum over a finite grid of priors of resolution η, at a cost of η on the value. Combined with the truncated-belief algorithm of Section 4.3, this yields:

[4] Mixed strategies are (finite support) distributions over randomized strategies.

▶ **Theorem 22** (adapted from Theorem 10 of [2]). *The gap problem for the universal value of a parity objective on an MEMDP is in EXPSPACE; if the probabilities are given in unary, it is in PSPACE.*

This improves on the doubly-exponential-space construction of [8] sketched in Section 3.3: the prior-value gap problem is solved in PSPACE (when probabilities are given in unary) thanks to the truncated-belief construction, and Theorem 21 with Lipschitz continuity lifts the bound to the universal-value gap.

Lower bound. It can be shown that, under the prior semantics, the gap problem is already NP-hard even when restricted to two environments and unary encoding of probabilities. This follows from an NP-hardness result established in [14] for the universal semantics; the corresponding details are provided in [2]. However, additional work is required to precisely determine the complexity of the gap problem.

4.5 Entropy and Dirac-preserving POMDPs

Both the truncated-belief approach and the connection between universal and prior semantics depend on a feature particular to MEMDPs equipped with a prior: once the belief over the hidden component becomes Dirac, it remains Dirac. This feature can also be formulated in the broader framework of POMDPs.

▶ **Definition 23** (Dirac-preserving POMDP, [2]). *A POMDP $\mathcal{P} = (S, A, \delta, \mathcal{O}, \mathrm{obs}, \iota_0)$ is* Dirac-preserving *if, for every state $s \in S$ (viewed as the Dirac belief $\mathbf{1}_s$), every action a and every observation o compatible with (s, a), the belief update at $(\mathbf{1}_s, a, o)$ is again Dirac.*

In words, once the controller knows the state for sure, it keeps knowing it for sure and for ever. By Proposition 19, every MEMDP with a prior is Dirac-preserving.

An equivalent characterisation uses entropy. For $b \in \mathcal{D}(X)$, the Shannon entropy is $H(b) = -\sum_{x \in X} b(x) \log b(x) \geq 0$; it is 0 for a Dirac distribution and $\log |X|$ for a uniform one. Following [5], a POMDP has *non-increasing expected entropy* if, for every belief $b \in \mathcal{D}(S)$ and every action a, $H(b) \geq \sum_{o \in \mathrm{Comp}(b,a)} p(b, a, o) \cdot H\big(\mathsf{Update}(b, a, o)\big)$, where $\mathrm{Comp}(b, a)$ is the set of observations possible under (b, a), $p(b, a, o)$ is their likelihood, and $\mathsf{Update}(b, a, o)$ is the corresponding posterior.

Non-increasing expected entropy clearly implies Dirac-preservation (otherwise entropy could rise from 0). The converse is more delicate, and the details can be found in [2].

▶ **Proposition 24** (Proposition 14 of [2]). *A POMDP is Dirac-preserving iff it has non-increasing expected entropy.*

The natural information-theoretic class thus matches a syntactic class whose membership can be decided in polynomial time (by checking all (s, a, o) triples). This class strictly includes MEMDPs with a prior, yet it remains tractable from an algorithmic perspective, since one can show that a Dirac-preserving POMDP can be converted into a (potentially exponentially larger) prior MEMDP with the same value for a general class of objectives.

▶ **Theorem 25** (Theorem 15 of [2]). *Let $\mathcal{P}$ be a Dirac-preserving POMDP with an observation-compatible parity objective Ω and an initial belief b. One can compute in exponential time an MEMDP $\mathcal{E}$, a parity objective Ω' and a prior μ such that $\mathsf{val}(\mathcal{P}, b, \Omega) = \mathsf{val}^P(\mathcal{E}, \mu, \Omega')$.*

Therefore, modulo an exponential blow-up in size, MEMDPs with a prior correspond exactly to Dirac-preserving POMDPs. Given that the limit-sure and gap problems are undecidable for general POMDPs – even when restricted to observation-compatible parity objectives – Dirac-preservation is exactly the structural condition that accounts for the decidability of these problems in MEMDPs endowed with a prior.

In [10], the authors independently defined a subclass of POMDPs, called Posterior-Deterministic POMDPs, which is exactly the same class as the Dirac-preserving POMDPs we studied in [2].

5 Other Related Works

Efficient belief computation. [5] addresses the practical side of the prior-semantics value problem. In an MEMDP, the belief over environments updates in time $O(|\mathcal{I}|)$ per step, against $O(|S|^2)$ for a general POMDP.

Multi-environment POMDPs. A natural extension of MEMDP is to combine MEMDP-style static uncertainty with genuinely partial observability of the state. Bovy, Probine, Suilen, Topcu and Jansen [3] study *multi-environment POMDPs* (MEPOMDPs) and show that, in the worst-case semantics, their analysis reduces to that of a single POMDP whose hidden state space is the product of the original states and the environment index, inheriting the undecidability of general POMDPs. Brice, Cano, Chatterjee, Henzinger and Muroya [4] focus instead on MEPOMDPs with *finite-horizon* reward objectives: they prove the value problem PSPACE-complete and provide a space-efficient deterministic algorithm that empirically outperforms the previous best.

Contextual MDPs and Multi-model MDPs. The MEMDP model has been rediscovered independently in adjacent communities under different names. Two notable examples are *Contextual MDPs* (CMDPs) [11] in the AI and reinforcement-learning community, and *Multi-model MDPs* (MMDPs) [15] in the operations-research community. In both, a hidden static index selects one MDP among a finite family sharing the same state and action spaces. The underlying mathematical object is essentially that of an MEMDP. The questions asked, however, are different. CMDPs are studied in episodic reinforcement learning with a fixed finite horizon T. At every episode the learner interacts with a trajectory drawn from a latent context, and the objective is to minimise *regret*: simultaneously learn the family of models and identify the active context. The focus is statistical, with sample-complexity and regret guarantees. MMDPs are studied in operations research, motivated by applications in medical decision-making. The typical objective is to optimise a weighted average of the value across all models; algorithmic tools include MILP formulations for the finite-horizon problem and policy-based branch-and-bound for the infinite-horizon one. The MEMDP line, in contrast, focuses on verification and synthesis for infinite-horizon ω-regular objectives (reachability, parity, Rabin), under universal or prior semantics, and asks for decidability, exact complexity and the structure of optimal strategies, rather than regret bounds or weighted-sum optimisation.

Robust MDPs. An other closely related model is that of *Robust MDPs* (RMDPs) [13, 12, 19]. An RMDP equips each pair (s, a) with an *uncertainty set* $\mathcal{P}(s, a) \subseteq \mathcal{D}(S)$; the value of a strategy is the worst case when, *at every step*, an adversary picks the actual distribution from $\mathcal{P}(s, a)$. The key difference with MEMDPs is the temporal resolution of uncertainty, it is

static in MEMDPs (one environment, drawn once), while it is dynamic in RMDPs (adversary picks at each step). Under the universal semantics MEMDPs let the controller learn the environment, whereas RMDPs do not. The two models are therefore formally incomparable.

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. Multi-environment pomdps: Discrete model uncertainty under partial observability. *CoRR*, abs/2510.23744, 2025. doi:10.48550/arXiv.2510.23744.

4 Léonard Brice, Filip Cano, Krishnendu Chatterjee, Thomas A. Henzinger, and Stefanie Muroya. Multi-environment POMDPs with finite-horizon objectives. *CoRR*, abs/2605.07537, 2026. doi:10.48550/arXiv.2605.07537.

5 Krishnendu Chatterjee, Martin Chmelík, Deep Karkhanis, Petr Novotný, and Amélie Royer. Multiple-environment Markov decision processes: Efficient analysis and applications. *Proc. Int. Conf. Autom. Plan. Sched.*, 30(1):48–56, 2020. doi:10.1609/ICAPS.V30I1.6644.

6 Krishnendu Chatterjee, Martin Chmelík, and Mathieu Tracol. What is decidable about partially observable Markov decision processes with ω-regular objectives. *Comput. Sci. J. Moldova*, 24(2):214–239, 2016. URL: https://csjournal.rm.

7 Krishnendu Chatterjee, Laurent Doyen, and Thomas A. Henzinger. Qualitative analysis of partially-observable markov decision processes. In *MFCS*, Lecture Notes in Computer Science, pages 258–269. Springer, 2010. doi:10.1007/978-3-642-15155-2_24.

8 Krishnendu Chatterjee, Laurent Doyen, Jean-François Raskin, and Ocan Sankur. The value problem for multiple-environment MDPs with parity objective. In *52nd International Colloquium on Automata, Languages, and Programming, ICALP 2025*, volume 334 of *LIPIcs*, pages 150:1–150:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPIcs.ICALP.2025.150.

9 Luca de Alfaro. *Formal verification of probabilistic systems*. PhD thesis, Stanford University, USA, 1997. URL: https://searchworks.stanford.edu/view/3910936.

10 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.

11 Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *CoRR*, abs/1502.02259, 2015. arXiv:1502.02259.

12 Garud N. Iyengar. Robust dynamic programming. *Math. Oper. Res.*, 30(2):257–280, 2005. doi:10.1287/MOOR.1040.0129.

13 Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.*, 53(5):780–798, 2005. doi:10.1287/OPRE.1050.0216.

14 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*, volume 29 of *LIPIcs*, pages 531–543. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2014. doi:10.4230/LIPIcs.FSTTCS.2014.531.

15 Lauren N. Steimle, David L. Kaufman, and Brian T. Denton. Multi-model markov decision processes. *IISE Trans.*, 53(10):1124–1139, 2021. doi:10.1080/24725854.2021.1895454.

16 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*, volume 311 of *LIPIcs*, pages 40:1–40:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi:10.4230/LIPIcs.CONCUR.2024.40.

17 Marck van der Vegt, Nils Jansen, and Sebastian Junges. Robust almost-sure reachability in multi-environment MDPs. In Sriram Sankaranarayanan and Natasha Sharygina, editors, *Tools and Algorithms for the Construction and Analysis of Systems - 29th International Conference, TACAS 2023*, volume 13993 of *Lecture Notes in Computer Science*, pages 508–526. Springer, 2023. doi:10.1007/978-3-031-30823-9_26.

18 Moshe Y. Vardi. Automatic verification of probabilistic concurrent finite-state programs. In *FOCS*, pages 327–338. IEEE Computer Society, 1985. doi:10.1109/SFCS.1985.12.

19 Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Math. Oper. Res.*, 38(1):153–183, 2013. doi:10.1287/MOOR.1120.0566.