Multiple-Environment Markov Decision Processes*

Jean-François Raskin and Ocan Sankur

Département d'Informatique, Université Libre de Bruxelles (U.L.B.), Belgium

Abstract

We introduce Multi-Environment Markov Decision Processes (MEMDPs) which are MDPs with a *set* of probabilistic transition functions. The goal in an MEMDP is to synthesize a single controller strategy with guaranteed performances against *all* environments even though the environment is unknown a priori. While MEMDPs can be seen as a special class of partially observable MDPs, we show that several verification problems that are undecidable for partially observable MDPs, are decidable for MEMDPs and sometimes have even efficient solutions.

1998 ACM Subject Classification D.2.4 Software/Program verification, F.3.1 Specifying and Verifying and Reasoning about Programs, G.3 Probabilities and Statistics

Keywords and phrases Markov decision processes, probabilistic systems, multiple objectives

Digital Object Identifier 10.4230/LIPIcs.FSTTCS.2014.531

1 Introduction

Markov decision processes (MDP) are a standard formalism for modeling systems that exhibit both stochastic and non-deterministic aspects. At each round of the execution of an MDP, an action is chosen by a controller (resolving non-determinism), and the next state is determined stochastically by a probability distribution associated to the current state and the chosen action. A controller is thus a *strategy* (a.k.a. *policy*) that determines which action to choose at each round according to the history of the execution so far. Algorithms for finite state MDPs are known for a large variety of objectives including omega-regular objectives [5], PCTL objectives [1], or quantitative objectives [18].

Multiple-Environment MDP (MEMDP). In an MDP, the environment is *unique*, and this may not be realistic: we may want to design a control strategy that exhibits good performances under several hypotheses formalized by different models for the environment, and those environments may *not* be distinguishable or we may *not want* to distinguish them (e. g. because it is too costly to design several control strategies.) As an illustration, consider the design of guidelines for a medical treatment that needs to work adequately for two populations of patients modeled by different stochastic models, even if the patients cannot be diagnosed to be in one population or in the other. This can be modeled by an MDP with two different models for the responses of the patients to the sequence of actions taken during the cure. We want a therapy that possibly makes decisions by observing the reaction of the patient and that works well (say reaches a good state for the patient with high probability) no matter if the patient belongs to the first of the second population.

Facing two potentially indistinguishable environments can be easily modeled with a partially observable MDPs. Unfortunately, this model is particularly intractable [3] (e. g. quantitative reachability, safety, and parity objectives, and even qualitative parity objectives

* Supported by the ERC starting grant inVEST (FP7-279499).

are undecidable.) To remedy to this situation, we introduce *multiple-environment MDPs (MEMDP)* which are MDPs with a *set* of probabilistic transition functions, rather than a *single one*. The goal in a MEMDP is to synthesize a single controller with guaranteed performances against *all* environments even though the environment which is operating is unknown a priori (it may be discovered during interaction but not necessarily.) We show that problems that are undecidable for partially observable MDPs, are decidable for MEMDPs and sometimes have even efficient solutions.

Results. We study MEMDPs with three types of objectives: reachability, safety and parity objectives. For each of those objectives, we study both *qualitative* and *quantitative* threshold decision problems. In this paper, we concentrate on MEMDPs with two environments as the two-environment case exhibits all the conceptual difficulties of the general case and it will easy the presentation of our results. The generalisation of the results for the n-environment case is left for future works. We first show that winning strategies may need infinite memory as well as randomization, and we provide algorithms to solve the decision problems. As it is classical, we consider two variants for the qualitative threshold problems. The first variant, asks to determine the existence of a *single* strategy that wins the objective with probability one (almost surely winning) in all the environments of the MEMDP. The second variant asks to determine the existence of a family of *single* strategies such that for all $\epsilon > 0$, there is one strategy in the family that wins the objective with probability larger than $1 - \epsilon$ (limit sure winning) in all the environments of the MEMDP. For both almost sure winning and limit sure winning, and for all three types of objectives, we provide efficient polynomial time algorithmic solutions. Then we turn to the quantitative threshold problem that asks for the existence of a single strategy that wins the objective with a probability that exceeds a given rational threshold in all the environments. We show the problem to be NP-hard (already for two environments and acyclic MEMDPs), so classical quantitative analysis techniques based on LP cannot be applied easily. Instead, we show that finite memory strategies are sufficient to approach achievable thresholds and we reduce the existence of bounded memory strategies to solving quadratic equations, leading to solutions in polynomial space. Our solutions rely on several new concepts (double-end components, good end-component, revealing edges, etc.) that bring deep understanding of the problems. The proofs are omitted due to space constraints, but a long version is available in [19].

Related Work. Interval Markov chains are Markov chains in which transition probabilities are only known to belong to given intervals (see *e. g.* [13, 14, 4]). Similarly, Markov decision processes with uncertain transition matrices for finite-horizon and discounted cases were considered [17]. The latter work also mentions the *finite scenario case* in which the transition probabilities are chosen among a finite set, as in our case. However, a solution is given only for the case where these probabilities can *independently* change in each step. Independence is a *simplifying assumption* that provides pessimistic guarantees. This means one ignores all information one might obtain on the system along observed histories, and so the results tend to be overly pessimistic.

Our work is related to reinforcement learning, where the goal is to develop strategies which ensure good performance in unknown environments, by learning and optimizing simultaneously; see [12] for a survey. In particular, it is related to the multi-armed bandit problem where one is given a set of systems with unknown reward distributions, and the goal is to choose the best one while optimizing the overall cost incurred while learning. The problem of finding the optimal one (without optimizing) with high confidence was considered

in [9, 15], and is related to our constructions inside *distinguishing double end-components* (see Section 5). However, our problems differ from this one as in multi-armed bandit problem models of the bandits are unknown while our environments are known but we do not know a priori which environment is playing.

Multiple reachability objectives in MDPs were considered in [7]: given an MDP and multiple targets T_i, thresholds α_i, decide if there is a strategy forcing each T_i with probability at least α_i. Multiple reachability in MDPs can be seen as a special case of the reachability problem in MEMDPs (consider multiple copies of the same transition relation and for each copy one target set T_i) but there is no easy reduction the other way around. Indeed, while for multi-reachability objectives in MDPs with absorbing target states, optimal memoryless strategies always exist [7], we show that for reachability objectives in MEMDPs with absorbing target states, we may need infinite memory to play optimally. The former problem can be solved in polynomial time using linear programming; but we show that the quantitative reachability problem for MEMDPs with *two* environments and absorbing target states is NP-hard; so no polynomial time reduction to multi-reachability in MDPs is possible unless P=NP. An extension to multiple quantitative objectives were considered in [10], where finite-memory strategies also suffice and the algorithm is uses linear programming.

2 Definitions

A finite *Markov decision process (MDP)* is a tuple $M = (S, A, \delta)$, where S is a finite set of *states*, A a finite set of *actions*, where $A(s)$ denotes the set of actions available from $s \in S$, and $\delta : S \times A \to \mathcal{D}(S)$ a partial function defined for each pair (s, a) such that $a \in A(s)$, and $\mathcal{D}(S)$ is the set of *probability distributions* on S. We define a *run* of M as a finite or infinite sequence $s_1 a_1 \dots a_{n-1} s_n \dots$ of states and actions such that $\delta(s_i, a_i, s_{i+1}) > 0$ for all $i \geq 1$. Finite runs are also called *histories* and denoted $\mathcal{H}(M)$.

Sub-MDPs and end-components. For the following definitions, we fix an MDP $M = (S, A, \delta)$. A *sub-MDP* M' *of* M is an MDP (S', A', δ') with $S' \subseteq S$, $\emptyset \neq A'(s) \subseteq A(s)$ for all $s \in S'$, and $\mathsf{Supp}(\delta(s, a)) \subseteq S'$ for all $s \in S', a \in A'(s)$ (here $\mathsf{Supp}(\cdot)$ denotes the support), and $\delta' = \delta|_{S' \times A'}$. By an abuse of notation, we may omit δ', and refer to the sub-MDP by (S', A'). For any subset $S' \subseteq S$ for which there exists a sub-MDP (S', A', δ'), let us denote by $M|_{S'}$ the *sub-MDP of* M *induced by* S', which is the sub-MDP with the largest set of actions. In other terms, the sub-MDP induced by S' contains *all* actions of S' whose supports are inside S'. An MDP is strongly connected if between any pair of states s, t, there is a run. An *end-component* of $M = (S, A, \delta)$ is a sub-MDP $M' = (S', A', \delta')$ that is strongly connected. It is known that the union of two end components with non-empty intersection is an end-component; one can thus define *maximal* end-components. We let $\mathsf{MEC}(M)$ denote the set of maximal end-components of M, computable in polynomial time [6]. An *absorbing state* s is such that for all $a \in A(s)$, $\delta(s, a, s) = 1$.

Strategies. A *strategy* σ is a function $(SA)^*S \to \mathcal{D}(A)$ such that for all $h \in (SA)^*S$ ending in s, we have $\mathsf{Supp}(\sigma(h)) \subseteq A(s)$. A strategy is *pure* if all histories are mapped to *Dirac distributions*. A strategy σ can be encoded by a *stochastic Moore machine*, $(\mathcal{M}, \sigma_a, \sigma_u, \alpha)$ where $\mathcal{M}$ is a finite or infinite set of memory elements, α the *initial distribution on* $\mathcal{M}$, σ_u the *memory update function* $\sigma_u : A \times S \times \mathcal{M} \to \mathcal{D}(\mathcal{M})$, and $\sigma_a : S \times \mathcal{M} \to \mathcal{D}(A)$ the *next action function* where $\mathsf{Supp}(\sigma(s, m)) \subseteq A(s)$ for any $s \in S$ and $m \in \mathcal{M}$. We say that σ is *finite-memory* if $|\mathcal{M}| < \infty$, and K*-memory strategy* if $|\mathcal{M}| = K$; it is memoryless

if $K = 1$, thus only depends on the last state of the history. Otherwise a strategy is *infinite-memory*. We define such strategies as functions $s \mapsto \mathcal{D}(A(s))$ for $s \in S$. An MDP M, a strategy σ encoded by $(\mathcal{M}, \sigma_a, \sigma_u, \alpha)$, and a state s determine a finite Markov chain M_s^σ defined on the state space $S \times \mathcal{M}$ as follows. The initial distribution is such that for any $m \in \mathcal{M}$, state (s, m) has probability $\alpha(m)$, and 0 for other states. For any pair of states (s, m) and (s', m'), the probability of the transition $(s, m), a, (s', m')$ is equal to $\sigma_a(s, m)(a) \cdot \delta(s, a, s') \cdot \sigma_u(s, m, a)(m')$. A *run* of M_s^σ is a finite or infinite sequence of the form $(s_1, m_1), a_1, (s_2, m_2), a_2, \ldots$, where each $(s_i, m_i), a_i, (s_{i+1}, m_{i+1})$ is a transition with nonzero probability in M_s^σ, and $s_1 = s$. In this case, the run $s_1 a_1 s_2 a_2 \ldots$, obtained by projection to M, is said to be *compatible with* σ. When considering the probabilities of events in M_s^σ, we will often consider sets of runs of M. Thus, given $E \subseteq (SA)^*$, we denote by $\mathbb{P}^\sigma_{M,s}[E]$ the probability of the runs of M_s^σ whose projection to M is in E. For any strategy σ in a MDP M, and a sub-MDP $M' = (S', A', \delta')$, we say that σ is *compatible with* M' if for any $h \in (SA)^* S'$, $\mathsf{Supp}(\sigma(h)) \subseteq A'(\mathsf{last}(h))$, where $\mathsf{last}(h)$ is the last state of h.

Let $\mathsf{Inf}(w)$ denote the disjoint union of states and actions that occur infinitely often in the run w; Inf is thus seen as a random variable. By an abuse of notation, we say that $\mathsf{Inf}(w)$ is equal to a sub-MDP D whenever it contains exactly the states and actions of D. It was shown that for any MDP M, state s, strategy σ, $\mathbb{P}^\sigma_{M,s}[\mathsf{Inf} \text{ is an end-component}] = 1$ [6]. We call a subset of states T *transient* if under all strategies, and starting from any state, almost surely, T is visited finitely many times.

Objectives. Given a set T of states, we define a *safety objective w.r.t.* T, written $\mathsf{Safe}(T)$, as the set of runs that only visit T. A *reachability objective w.r.t.* T, written $\mathsf{Reach}(T)$, is the set of runs that visit T at least once. We also consider *parity objectives*. A *parity function* is defined on the set of states $p : S \to \{0, 1, \ldots, 2d\}$ for some nonnegative integer d. The set of *winning runs of* M *for* p is defined as $\mathcal{P}_p = \{w \in (SA)^\omega \mid \min\{p(s) \mid s \in \mathsf{Inf}(w)\} \in 2\mathbb{N}\}$. For any MDP M, state s, strategy σ, and objective Φ, we denote $\mathsf{Val}^\sigma_\Phi(M, s) = \mathbb{P}^\sigma_{M,s}[\Phi]$ and $\mathsf{Val}^*_\Phi(M, s) = \sup_\sigma \mathbb{P}^\sigma_{M,s}[\Phi]$. We say that objective Φ is *achieved surely* if for some σ, all runs of M from s compatible with σ satisfy Φ. Objective Φ is *achieved with probability* α in M from s if for some σ, $\mathsf{Val}^\sigma_\Phi(M, s) \geq \alpha$. If Φ is achieved with probability 1, we say that it is *achieved almost surely*. Objective Φ is *achieved limit-surely* if for any $\epsilon > 0$, there exists a strategy σ_ϵ which achieves Φ with probability $1 - \epsilon$. In MDPs, limit-sure achievability coincides with almost-sure achievability since optimal strategies exist. We define $\mathsf{AS}(M, \Phi)$ as the set of states of M where Φ is achieved almost surely. Recall that for reachability, safety, and parity objectives these states can be computed in polynomial time, and are only dependent on the supports of the probability distributions [1, 6]. It is known that for any MDP M, state s, and a reachability, safety, or parity objective, there exists a pure memoryless strategy σ computable in polynomial time achieving the optimal value [18, 5].

In the next lemma, we recall that the optimal value inside any end-component is either 0 or 1, and that this only depends on the supports of the probability distributions.

▶ **Lemma 1** ([6]). *Let $M = (S, A, \delta)$ be a strongly connected MDP, and p a parity function. Then, for any MDP $M' = (S, A, \delta')$ such that for all $s \in S$, $a \in A$, $\mathsf{Supp}(\delta(s, a)) = \mathsf{Supp}(\delta'(s, a))$, and for all states $s \in S$, there exists a strategy σ such that $\mathsf{Val}^\sigma_{\mathcal{P}_p}(M, s) = \mathsf{Val}^*_{\mathcal{P}_p}(M, s) = \mathsf{Val}^*_{\mathcal{P}_p}(M', s) = \mathsf{Val}^\sigma_{\mathcal{P}_p}(M', s) \in \{0, 1\}$.*

3 Multiple-Environment MDP

A *multiple-environment MDP (MEMDP)*, is a tuple $M = (S, A, (\delta_i)_{1 \leq i \leq k})$, where for each i, (S, A, δ_i) is an MDP. We will denote by M_i the MDP obtained by fixing the edge proba-

bilities δ_i, so that $\mathbb{P}^{\sigma}_{M_i,s}[E]$ denotes the probability of event E in M_i from state s under strategy σ. Intuitively, each M_i corresponds to the behavior of the system at hand under a different *environment*; in fact, while the state space is identical in each M_i, the transition probabilities between states and even their supports may differ. In this paper, we concentrate on the case of $k = 2$. We are interested in synthesizing a *single* strategy σ with guarantees on *both* environments, without a priori knowing against which environment σ is playing. We consider reachability, safety, and parity objectives, and again for readability, we consider the case where the same objective is to hold in all environments. The general quantitative problem is the following.

▶ **Definition 2.** Given MEMDP M, state s_0, $\alpha_1, \alpha_2 \in [0,1]$, and Φ, a reachability, safety, or a parity objective, decide if there is a strategy σ such that $\forall i \in \{1,2\}, \mathsf{Val}^{\sigma}_{\Phi}(M_i, s) \geq \alpha_i$.

We refer to the general problem as *quantitative reachability (resp. safety, parity)*. Given M, s_0, (α_1, α_2), Φ, we say that Φ is *achieved with probabilities* (α_1, α_2) in M from s if there is a strategy σ witnessing the above definition. We say that Φ is *achieved almost surely* in M from s if it is achieved with probabilities $(1,1)$. Objective Φ is *achieved limit-surely* in M from s if for any $\epsilon > 0$, Φ is achieved in M from s with probabilities $(1-\epsilon, 1-\epsilon)$. *Almost-sure reachability (resp. safety, parity)* problems consist in deciding whether in a given M, from a state s, a given objective is achieved almost surely. *Limit-sure reachability (resp. safety, parity)* problems are defined respectively.

Given any MEMDP $M = (S, A, \delta_1, \delta_2)$, we define the MDP $\cup M = (S, A, \delta)$ by taking, for each action, the union of all transitions, and assigning them uniform probabilities. For any sub-MDP (S', A', δ') of $\cup M$, we define the *sub-MEMDP induced by the sub-MDP* (S', A', δ') as the MEMDP $(S', A', \delta'_1, \delta'_2)$ where $\delta'_i = \delta_i|_{S' \times A'}$. For any subset $S' \subseteq S$, the *sub-MEMDP of M induced by S'* is the sub-MEMDP of M induced by the sub-MDP of $\cup M$ induced by S'.

Strategy Complexity. Unlike MDPs, all considered objectives may require infinite memory and randomization, and Pareto-optimal probability vectors may not be achievable (a Pareto-optimal vector is component-wise maximal). All counterexamples are given in Fig. 1.

▶ **Lemma 3.** *For some MEMDPs M and objectives Φ:*

- *There exists a randomized strategy that achieves Φ with higher probabilities in both environments than any pure strategy,*
- *There exists an infinite-memory strategy that achieves Φ with higher probabilities in both environments than any finite-memory strategy,*
- *Objective Φ can be achieved limit-surely but not almost surely (showing Pareto-optimal vectors are not always achievable).*

Results. We give efficient algorithms for almost-sure and limit-sure cases:

(A) The almost-sure reachability, safety, and parity problems are decidable in polynomial time (Theorems 5 and 19). Finite-memory strategies suffice.

(B) The limit-sure reachability, safety, and parity problems are decidable in polynomial time (Theorems 12 and 20). Moreover, for any $\epsilon > 0$, to achieve probabilities of at least $1-\epsilon$, $O(\frac{1}{\eta^2}\log(\frac{1}{\epsilon}))$-memory strategies suffice, where η denotes the smallest positive difference between the probabilities of M_1 and M_2.

The general quantitative problem is harder as shown by the next results. We call an MEMDP *acyclic* if the only cycles are self-loops in all environments.

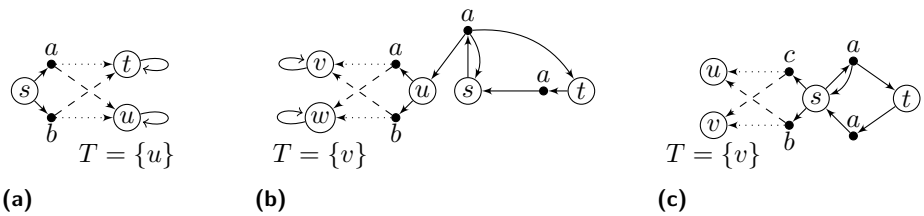


Figure 1 We adopt the following notation in all examples: edges that only exist in M_1 are drawn in dashed lines, and those that only exist in M_2 by dotted ones, and all probabilities are uniform unless otherwise said. To see that randomization may be necessary, observe that in the MEMDP M in Fig. 1a, the vector $(0.5, 0.5)$ of reachability probabilities for target T can only be achieved by a strategy that randomizes between a and b. In the MEMDP in Fig. 1b, where action a from s has the same support in M_1 and M_2 but different distributions. Any strategy almost surely reaches u in both M_i, since action a from s has nonzero probability of leading to u. Intuitively, the best strategy is to sample the distribution of action a from s, and to choose, upon arrival to u, either a or b according to the most probable environment. We prove that such an infinite-memory strategy achieves a Pareto-optimal vector which cannot be achieved by any finite-memory strategy. Last, in Fig. 1c, the MEMDP is similar to that of Fig. 1b except that action a from s only leads to s or t. We will prove in Section 6, that for any $\epsilon > 0$, there exists a strategy ensuring reaching T with probability $1 - \epsilon$ in each M_i. The strategy consists in sampling the distribution of action a from s a sufficient number of times and guessing the actual environment against which the controller is playing. However, the vector $(1, 1)$ is not achievable, which follows from Section 4.

(C) The quantitative reachability and safety problems are NP-hard on acyclic MEMDPs both for arbitrary and memoryless strategies (Theorem 13).

We can nevertheless provide procedures to solve the quantitative reachability and safety problems by fixing the memory size of the strategies.

(D) For any $K \geq 0$, the quantitative reachability and safety problems restricted to K-memory strategies can be solved in space polynomial in K and the size of M. (Theorem 14).

The quantitative parity problem can be reduced to quantitative reachability, so the previous result can also be applied for the quantitative parity problem.

(E) The quantitative parity problem can be reduced to quantitative reachability in polynomial time (Theorem 20).

We show that finite-memory strategies are not restrictive if we are interested in approximately ensuring given probabilities.

(F) Finite-memory strategies suffice to approximate quantitative reachability, safety, and parity problems up to any desired precision (Theorem 15).

We provide an *approximate* solution for quantitative reachability in the following sense. We consider *ϵ-gap problems* where the goal is to give a correct answer on negative instances that are "far" from the positive instances by ϵ, and on positive instances that are far from the negative instances by ϵ, while giving no guarantees in the rest of the input [8, 11].

▶ **Definition 4.** The *ϵ-gap problem for reachability* consists, given MEMDP M, state s, target set T, and probabilities α_1, α_2, in answering i YES if $\exists\sigma, \forall i = 1, 2, \mathbb{P}^{\sigma}_{M_i,s}[\mathsf{Reach}(T)] \geq \alpha_i$, ii NO if $\forall\sigma, \exists i = 1, 2, \mathbb{P}^{\sigma}_{M_i,s}[\mathsf{Reach}(T)] < \alpha_i - \epsilon$, iii and arbitrarily otherwise.

(G) There is a procedure for the ϵ-gap problem for quantitative reachability in MEMDPs (Theorem 16). The ϵ-gap problem is NP-hard (Theorem 17).

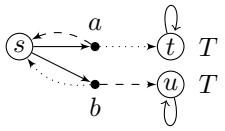


Figure 2 MEMDP M where $\mathsf{Reach}(T)$ can be achieved almost surely. In fact, $\mathsf{AS}(M_i, T) = \{s, t, u\}$ for all $i = 1, 2$, so $M' = M$, and $\mathsf{Val}_{\mathsf{Reach}(T)}(M'_i, s) = 1$ for $i = 1, 2$. The strategy returned by the algorithm consists in choosing, at s, a and b uniformly at random. Notice that there is no pure memoryless strategy achieving the objective almost surely.

Preprocessing. In an MEMDP with two environments, if one observes an edge that only exists in one environment, then the environment is known with certainty and any good strategy should immediately switch to the optimal strategy for the revealed environment. Formally, an edge (s, a, s') is *i-revealing* if $\delta_i(s, a, s') \neq 0$ and $\delta_{3-i}(s, a, s') = 0$. We make the following assumption w.l.o.g.:

▶ Assumption 1 (Revealed form). All MEMDPs $M = (S, A, \delta_1, \delta_2)$ are assumed to be in *revealed form*, that is, there exists a partition $S = S_u \uplus R_1 \uplus R_2$ satisfying the following properties. **1.** All states of R_1 and R_2 are absorbing in both environments, **2.** For any $i = 1, 2$, and any i-revealing edge (s, a, s'), we have $s' \in R_i$. Conversely, any edge (s, a, s') with $s' \in R_i$ is i-revealing. States R_i are called *i-revealed*, denoted $R_i(M)$. We write $R = R_1 \cup R_2$. The remaining states are called *unrevealed*.

In other words, we assume that any i-revealing edge leads to a known set of i-revealed states which are all absorbing. Assumption 1 can be made without loss of generality by redirecting any revealing edge to fresh absorbing states.

For any reachability (resp. safety) objective T, once a state in T (resp. $S \setminus T$) is visited the strategy afterwards is not significant since the objective has already been fulfilled (resp. violated). Thus, we assume that the set of target and unsafe states are absorbing.

▶ Assumption 2. For all considered objectives $\mathsf{Reach}(T)$ (resp. $\mathsf{Safe}(T')$), we assume that all target states T (resp. unsafe states $S \setminus T'$) absorbing for both environments.

Under assumptions 1 and 2, for any MEMDP M, and objective Φ, we denote $R_i^\Phi(M)$ the set of i-revealed states from which Φ holds almost surely in M_i, and define $R^\Phi(M) = R_1^\Phi(M) \cup R_2^\Phi(M)$. We will apply Assumption 1 throughout the paper, and Assumption 2 for reachability or safety objectives.

4 Almost-Sure Reachability

The algorithm for almost sure reachability is described in Algorithm 1. First, the state space is restricted to U since any state from which the objective holds almost surely in the MEMDP M must also belong to an almost surely winning state of each M_i. Second, we consider MEMDP M' induced by the states surely satisfying $\mathsf{Safe}(U)$ in $\cup M$. The problem is then reduced to finding an almost surely winning strategy in each M'_i separately. If such strategies exist, then we obtain our strategy by either 1) alternating between the two strategies using memory, or 2) randomizing between them.

Figure 2 is an example where almost-sure reachability holds; and we saw the example of Fig. 1c where almost-sure reachability does not hold.

▶ **Theorem 5.** *For any MEMDP M, objective $\mathsf{Reach}(T)$, and a state s, Algorithm 1 decides in polynomial time if $\mathsf{Reach}(T)$ can be achieved almost surely from s in M, and returns a witnessing memoryless randomized strategy.*

Input: MEMDP M, $\mathsf{Reach}(T)$, $s_0 \in S$
$U := \mathsf{AS}(M_1, \mathsf{Reach}(T)) \cap \mathsf{AS}(M_2, \mathsf{Reach}(T))$;
$M' :=$ Sub-MEMDP of M induced by states s s.t. $\mathsf{Val}^*_{\mathsf{Safe}(U)}(\cup M, s) = 1$;
if $\forall i = 1, 2, \mathit{Val}^*_{\mathit{Reach}(T)}(M'_i, s_0) = 1$ **then**
  Let σ_i for $i = 1, 2$, such that $\mathsf{Val}^\sigma_{\mathsf{Reach}(T)}(M'_i, t) = 1$ for all $t \in U$;
  Return σ' defined as $\sigma'(t) = \frac{1}{2}\sigma_1(t) + \frac{1}{2}\sigma_2(t)$, $\forall t \in S$;
else
  Return NO;
end

Algorithm 1. Almost-sure reachability algorithm for MEMDPs.

5 Double end-components

End-components play an important role in the analysis of MDPs; see *e. g.* [6]. Because the probability distributions in different environments of an MEMDP can have different supports, we need to adapt the notion for MEMDPs. We thus introduce *double end-components* which are sub-MDPs that are end-components in both environments.

Formally, given an MEMDP $M = (S, A, \delta_1, \delta_2, r)$, a *double end-component (DEC)* is a pair (S', A') where $S' \subseteq S$, and $A' \subseteq A$ such that (S', A') is an end-component in each M_i. A double end-component (S', A') is *distinguishing* if there is $(s, a) \in S' \times A'$ such that $\delta_1(s, a) \neq \delta_2(s, a)$. As the union of two DECs with a common state is a DEC, we consider *maximal DECs (MDEC)*. MDECs of M can be computed in polynomial time by eliminating from M all actions with different supports, then computing the MECs of $\cup M$. A DEC is *trivial* if it is an absorbing state.

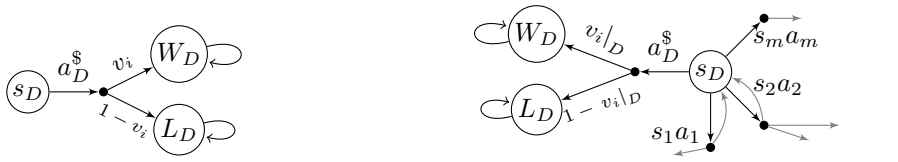
By Assumption 1 a DEC does not contain any revealed states unless it is trivial; therefore the supports of all DECs in both environments are identical. By Assumption 2, for reachability (resp. safety) objectives, non-trivial DECs do not contain target (resp. unsafe) states neither. Trivial DECs made of target (resp. safe) states are called *winning*. A DEC D is *winning* for a parity objective $\mathcal{P}_p$, if there is a strategy compatible with D satisfying $\mathcal{P}_p$ almost surely in both environments (a common strategy exists by Lemma 1).

Distinguishing DECs allow the strategy to learn the actual environment by sampling the distribution of distinguishing actions. One can in fact construct a strategy that surely stays inside a given DEC and guesses the actual environment with high confidence. Since a distinguishing DEC is non-trivial, the learning phase will surely avoid unsafe states. One then switches to the optimal strategy for the guessed environment:

▶ **Lemma 6.** *Consider any MEMDP $M = (S, s_0, A, \delta_1, \delta_2)$, a distinguishing double end-component $D = (S', A')$, state $s \in S'$, $\epsilon > 0$, and any objective Φ reachability, safety, parity. For any $\epsilon > 0$, there exists a strategy σ such that $\mathbb{P}^\sigma_{M_i,s}[\Phi] \geq (1 - \epsilon)\mathit{Val}^*_\Phi(M_i, s), \forall i = 1, 2$.*

We now present a transformation for general MEMDPs by contracting DECs, which preserves the values up to any desired ϵ by Lemma 6. Given a DEC $D = (S', A')$, a *frontier state* s of D is such that there exists an action $a \in A(s) \setminus A'(s)$, which is not in D, index $i \in \{1, 2\}$, and $s' \notin S'$ such that $\delta_i(s, a, s') \neq 0$. An action $a \in A(s) \setminus A'(s)$ is a *frontier action* for D. A pair (s, a) is called *frontier state-action* when $a \in A(s)$ is a frontier action.

▶ **Definition 7.** Given an MEMDP $M = (S, A, \delta_1, \delta_2)$, and reachability or safety objective Φ, we define $\hat{M} = (\hat{S}, \hat{A}, \hat{\delta_1}, \hat{\delta_2})$ as follows. a) Any distinguishing MDEC D is contracted as in Fig. 3a where in M_i, action a leads to W_D with probability $v_i = \mathsf{Val}^*_\Phi(M_i, D)$, and to



(a) Reducing distinguishing DECs, where $v_i = \mathsf{Val}^*_{\phi_i}(M_i, D)$.

(b) Reduction of double end-components where $v_i|_D = \mathsf{Val}^*_{\Phi}(M_i|_D, D)$.

Figure 3 Reduction of double end-components.

L_D with probability $1 - v_i$. b) Any non-distinguishing MDEC $D = (S', A')$ is replaced with the module in Fig. 3b. The actions $a_D^{\$}$ and $\{f_i a_i\}_{(f_i,a_i)\in F}$ are available from s_D where F is the set of pairs of frontier state-actions of D. For any (f_i, a_i), the distribution $\hat{\delta}_j(s_D, f_i a_i)$ is obtained from $\delta_j(f_i, a_i)$ by redirecting to s_D all edges that lead inside S'. Define the new objective $\hat{\Phi}$ by restricting Φ to $\hat{S}$, and adding all states W_D in the target (resp. safe) set. We write $\hat{\mathcal{A}} : S \to \hat{S}$ (also denoted $s \mapsto \hat{s}$) the mapping from the states of S to those of $\hat{S}$.

The intuition is that when the play enters a *distinguishing* DEC D, by Lemma 6, we can arbitrarily approximate probabilities $v_i = \mathsf{Val}^*_{\Phi}(M_i, D)$; this is represented by action $a_D^{\$}$ in Fig. 3a. From a state s in a *non-distinguishing* DEC D in M, the play either stays forever inside and obtains the value $\mathsf{Val}^*_{\Phi}(M_1|_D, s) = \mathsf{Val}^*_{\Phi}(M_2|_D, s)$ (as it is non-distinguishing) – represented by $a_D^{\$}$ in Fig. 3b, or it eventually leaves D. The latter case is represented by the actions leading to frontier states, since D is necessarily left from such a state. Note that there is a strategy under which, from any state of D, in M_1 and M_2, all states of D, and in particular its frontier states, are visited infinitely often (by considering a memoryless strategy choosing all actions uniformly at random – see *e. g.* [18]). The equivalence between M and $\hat{M}$ for reachability and safety is shown next. Note that the value vectors are preserved although vectors achieved in $\hat{M}$ may not be achievable in M (see Fig. 1c).

▶ **Lemma 8.** *For MEMDPs M, and reachability of safety objectives Φ, $\mathsf{Val}^*_{\Phi}(M, s) = \mathsf{Val}^*_{\hat{\Phi}}(\hat{M}, \hat{s})$. Any end-component D of $\hat{M}_i$ is either a trivial DEC, or transient in $\hat{M}_{3-i}$.*

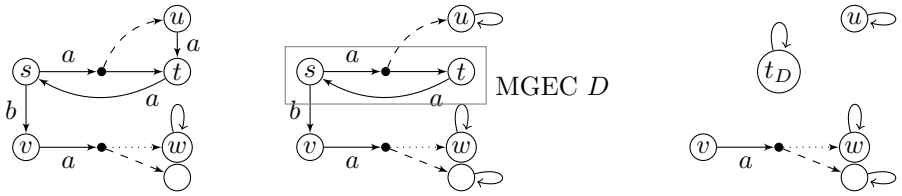
By Definition 7, and Lemmas 8-8, the following assumption can be made w.l.o.g.

▶ Assumption 3. All MEMDPs are assumed to have only trivial DECs.

6 Limit-Sure Reachability

In this section, we describe our polynomial-time algorithm for limit-sure reachability in MEMDPs. Throughout this section, we make Assumptions 1, 2, and 3.

We saw in the previous section how the strategy can safely learn the current environment with high confidence inside DECs. It turns out that it is possible to apply such learning strategies outside DECs. We need to introduce a new concept, called *good end-components* in order to fully capture all subsets of states where such a learning strategy can be applied. Consider the example of Fig. 4. Here, the MDP M_1 has a MEC D with the following property: the strategy σ compatible with D and choosing all actions of D uniformly at random, achieves the objective almost surely in M_2. In fact, a strategy that always chooses a at states s and t almost surely reaches u in M_2. In order to achieve the objective with probability close to 1, one can run strategy σ for a large number of steps, and if state u is still not reached, switch to the optimal strategy for M_1, that is, choose b from s. It can



(a) MEMDP M with $T = \{u, w\}$.

(b) Assumptions 1 and 2 satisfied. We have $T = \{u, w\}$.

(c) MEMDP $\widetilde{M}$ with $\widetilde{T} = \{t_D, u, w\}$ and $\widetilde{\mathcal{A}}(s) = t_D$.

Figure 4 On the left, an MEMDP with objective $\mathsf{Reach}(T)$, which is *not* in revealed form; an equivalent instance M in revealed form is shown in the middle. Note that M has only trivial DECs. States $\{s, t\}$ induce a *good end-component* D in M_2; the strategy choosing action a at s and t is almost surely winning in M_1. The construction $\widetilde{M}$ is shown on the right, where all states of D are contracted as t_D which is a target state. Because $\widetilde{\mathcal{A}}(s) = t_D$, Φ is achieved limit-surely from s.

be shown that such a strategy achieves the objective with probabilities $(1 - \epsilon, 1 - \epsilon)$, for any $\epsilon > 0$, from any state of such end-components. Here D is a *good end-component* of M_1.

Formally, an end-component D of M_i is *good* if the strategy that chooses all edges of D uniformly at random is almost sure winning for M_{3-i} from any state in D. Under any such strategy, any edge leaving D is revealing for M_{3-i}. Observe that the union of good end-components with non-empty intersection is a good end-component. We thus consider *maximal good end components (MGECs)* which can be computed in polynomial time.

▶ **Lemma 9.** *For any M, the MGECs of M_1 and M_2 can be computed in polynomial time.*

We define a transformation to MEMDPs by contracting MGECs since we know that one can learn the current environment from these states, without risking to lose.

▶ **Definition 10** ($\mathsf{Transformation}\ \widetilde{M}$)**.** For any MEMDP M, and reachability objective Φ, we let $\widetilde{M} = (\widetilde{S}, \widetilde{A}, \widetilde{\delta_1}, \widetilde{\delta_2})$ by applying the following transformation to M and Φ. Mark any state s that belongs to some MGEC D of M_i for some $i = 1, 2$, by D. If a state can be marked twice, choose one marking arbitrarily. We define $\widetilde{M}$ by redirecting any edge entering a state marked by some D to a fresh absorbing state t_D. For each $i = 1, 2$, the reachability objective $\widetilde{\Phi}$ is defined by the union of Φ, with all states t_D such that Φ can be ensured almost surely from D in M_i. We let $\widetilde{\mathcal{A}}(\cdot)$ be the mapping from the states M to those of $\widetilde{M}$.

Intuitively, DECs and MGECs cover all subsets of states in which one can learn the actual environment with high confidence; while in the absence of such components, limit-sure becomes equivalent to almost-sure. The following lemma establishes this property.

▶ **Lemma 11.** *For any MEMDP M, reachability objective Φ, and state s, Φ can be achieved limit surely in M from s if, and only if $\widetilde{\Phi}$ can be achieved almost surely in $\widetilde{M}$ from $\widetilde{\mathcal{A}}(s)$. Moreover, given an almost sure winning strategy for $\widetilde{M}$, for any $\epsilon > 0$, one can compute a strategy with memory $O(\frac{\log(\epsilon)}{\log(1-p)})$ for M, where p is the smallest nonzero probability, that achieves probabilities $1 - \epsilon$, and this strategy can be computed.*

The steps of the limit-sure reachability algorithm are thus as follows: **1.** Contract DECs by Def. 7. **2.** Contract MGECs by Def. 10. **3.** Solve almost-sure reachability by Algorithm 1.

▶ **Theorem 12.** *The limit-sure reachability problem is decidable in polynomial-time.*

7 Quantitative Reachability

In this section, we study the quantitative reachability problem for MEMDPs. We first establish NP-hardness, suggesting that it is unlikely to have a polynomial-time algorithm, and that techniques based on linear programming often used for the quantitative analysis of MDPs (e. g. [18]) cannot be applied. We prove the hardness result by reduction from the product-partition problem [16].

▶ **Theorem 13.** *Given an MEMDP M, target set T, and $\alpha_1, \alpha_2 \in [0,1]$, it is NP-hard to decide whether for some strategy σ, $\mathbb{P}^{\sigma}_{M_i,s_0}[\mathsf{Reach}(T)] \geq \alpha_i$ for each $i = 1, 2$.*

As an upper bound on the above problem, we show that quantitative reachability for strategies with a fixed memory size can be solved in polynomial space. The algorithm consists in encoding the strategy and the probabilities achieved by each state and each environment, as a bilinear equation, and solving these in polynomial space in the equation size (see [2] for general polynomial equations).

▶ **Theorem 14.** *The quantitative reachability and safety problems for K-memory strategies can be solved in polynomial space in K and in the size of M.*

We now show that considering finite-memory strategies are hardly restrictive, in the sense that they can always be used to approximately achieve the values. We give a bound on strategy memories that is sufficient to approximate the value by given ϵ. The idea underlying the proof of the following theorem is that along a long execution in MEMDPs, with high probability, either one enters a subset of states that is identical in both environments, or one has gathered enough samples on probability distributions to guess the actual environment with high confidence.

▶ **Theorem 15.** *For any MEMDP M satisfying Assumption 3, reachability objective Φ, strategy σ, and $\epsilon > 0$, there exists a N-memory strategy σ' with $\forall i = 1, 2, \mathbb{P}^{\sigma'}_{M_i,s}[\Phi] \geq \mathbb{P}^{\sigma}_{M_i,s}[\Phi] - \epsilon$, where $N = (|S| + |A|)^{\frac{4|S|^3|A|^2}{p^{|S|}\eta^2} \log^3(1/\epsilon)}$, with p the smallest nonzero probability and $\eta = \min\{|\delta_1(s,a,s') - \delta_2(s,a,s')| \mid s,a,s' \text{ s.t. } \delta_1(s,a,s') \neq \delta_2(s,a,s')\}$.*

We derive our procedure by Theorems 14 and 15. The "gap" can be chosen arbitrarily small, and the procedure is used to distinguish instances that are clearly feasible from those that are clearly not feasible, while giving no guarantee in the borderline of size ϵ.

▶ **Theorem 16.** *There is a procedure that runs in $O(N \cdot |M|)$ space solving the ϵ-gap problem for quantitative reachability in MEMDPs.*

It turns out that even the ϵ-gap problem is NP-hard. We prove this by identifying instances where the achieved probabilities are *isolated*:

▶ **Theorem 17.** *The ϵ-gap problem for MEMDPs is NP-hard.*

8 Safety and Parity Objectives

▶ **Lemma 18.** *Limit-sure safety is equivalent to sure safety in MEMDPs, and can be decided in polynomial time.*

For quantitative safety, the results of the previous section can be adapted without difficulty.

We give a polynomial-time algorithm for almost sure parity objectives, consisting in **1.** restricting the states to almost surely winning ones for both M_i, **2.** solving almost sure reachability where all states that belong to winning end-components in M_1 or M_2 are targets.

▶ **Theorem 19.** *The almost-sure parity problem is decidable in polynomial time.*

We now describe a polynomial-time reduction from quantitative parity to quantitative reachability preserving value vectors. The idea is to allow the strategy to irreversibly switch to an optimal strategy for environment i from any MEC of M_i, and to represent this switch by a target absorbing state. Intuitively, the new reachability condition is equivalent to the parity objective for two reasons: first, all runs eventually enter an end-component and stay there, which roughly corresponds to this switch, and second, the transformation only adds new actions, so any strategy in the original MEMDP is still valid in the new one, and in particular *learning* strategies. It follows 1) a polynomial-time algorithm for the limit-sure parity problem, 2) any algorithm for quantitative reachability can be used to solve the quantitative parity problem. In particular, results of Section 7 applies to parity.

▶ **Theorem 20.** *The quantitative parity problem is polynomial-time reducible to the quantitative reachability problem. The limit-sure parity problem is solvable in polynomial time.*

References

1 Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT Press, 2008.
2 John Canny. Some algebraic and geometric computations in pspace. In *STOC'88*, pages 460–467, New York, NY, USA, 1988. ACM.
3 Krishnendu Chatterjee, Martin Chmelik, and Mathieu Tracol. What is decidable about partially observable markov decision processes with omega-regular objectives. In *CSL*, volume 23 of *LIPIcs*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2013.
4 Taolue Chen, Tingting Han, and Marta Z. Kwiatkowska. On the complexity of model checking interval-valued discrete time markov chains. *Inf. Process. Lett.*, 113(7):210–216, 2013.
5 Costas Courcoubetis and Mihalis Yannakakis. The complexity of probabilistic verification. *J. ACM*, 42(4):857–907, July 1995.
6 Luca de Alfaro. *Formal verification of probabilistic systems*. Ph.D. thesis, Stanford University, 1997.
7 Kousha Etessami, Marta Z. Kwiatkowska, Moshe Y. Vardi, and Mihalis Yannakakis. Multi-objective model checking of Markov decision processes. *Logical Methods in Computer Science*, 4(4), 2008.
8 Shimon Even, Alan L. Selman, and Yacov Yacobi. The complexity of promise problems with applications to public-key cryptography. *Information and Control*, 61(2):159–173, 1984.
9 Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *COLT'02*, volume 2375 of *LNCS*, pages 255–270. Springer, 2002.
10 Vojtěch Forejt, Marta Kwiatkowska, Gethin Norman, David Parker, and Hongyang Qu. Quantitative multi-objective verification for probabilistic systems. In *TACAS'11*, volume 6605 of *LNCS*, pages 112–127. Springer, 2011.
11 Oded Goldreich. On promise problems (a survey in memory of Shimon Even [1935–2004]). *Manuscript*, 2005.
12 Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.
13 Igor O. Kozine and Lev V. Utkin. Interval-valued finite markov chains. *Reliable computing*, 8(2):97–113, 2002.
14 Antonín Kučera and Oldřich Stražovský. On the controller synthesis for finite-state markov decision processes. In *FSTTCS 2005*, volume 3821 of *LNCS*, pages 541–552. Springer, 2005.

15 Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648, December 2004.

16 C. T. Ng, M. S. Barketau, T. C. E. Cheng, and Mikhail Y. Kovalyov. "Product Partition" and related problems of scheduling and systems reliability: Computational complexity and approximation. *European Journal of Operational Research*, 207(2):601–604, 2010.

17 Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

18 Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

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