Multiple-Environment Markov Decision Processes

Jean-François Raskin Ocan Sankur

May 20, 2014

Abstract

We introduce Multi-Environment Markov Decision Processes (MEMDPs) which are MDPs with a *set* of probabilistic transition functions. The goal in a MEMDP is to synthesize a single controller 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.

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 a MDP, an action is chosen by a controller (resolving the non-determinism), and the next state is determined 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 [17].

Multiple-Environment MDP In a 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 (each given by a different stochastic model), even if the patients cannot be diagnosed to be in one population or in the other. A appropriate model for this case would be a 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 takes 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 modelled with a partially observable MDPs. Unfortunately, this model is particularly intractable [3] (e.g. qualitative and quantitative reachability, safety and parity objectives 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 at play is unknown a priori (it may be discovered during interaction but not necessarily.) We show that verification 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 [1]. 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), and so classical quantitative analysis techniques based on LP cannot be applied here. 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.

Related Work In addition to partially observable MDPs, our work is related to the following research lines.

Interval Markov chains are Markov chains in which transition probabilities are only known to belong to given intervals (see *e.g.* [12, 13, 4]). Similarly, Markov decision processes with uncertain transition matrices for finite-horizon and discounted cases were considered [16]. The latter work also mentions the finite scenario-case which is similar to our setting. However, the precise distributions of actions at each round are assumed to be independent while in our work we consider it to be fixed but unknown. Independence is a *simplifying assumption* that only provides pessimistic guarantees.

[1] For readability, we concentrate in this paper on MEMDPs with two environments, but most of the results can be easily generalized to any finite number of environments possibly with an increased computational complexity. This is left for a long version of this paper.

However this approach does not use the information one obtains 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 [11] for a survey. In particular, it is related to the multi-armed bandit problem where one is given a set of *stateless* systems with unknown reward distributions, and the goal is to choose the best one while optimizing the overall cost incured while learning. The problem of finding the optimal one (without optimizing) with high confidence was considered in [9, 14], 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 against which one we are playing.

MEMDPs are also related to multi-objective reachability in MDPs considered in [7], where a strategy is to be synthesized so as to ensure the reachability of a set of targets, each with a possibly different probability. If we allow multiple environments and possibly different reachability objectives for each environment, this problem can be reduced to reachability in MEMDPs. Note however that the general reachability problem is harder in MEMDPs; it is NP-hard even for acyclic MEMDPs with absorbing targets, while polynomial-time algorithms exist for absorbing targets in the setting of [7].

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*, and $\delta : S \times A \to \mathcal{D}(S)$ a partial function, where $\mathcal{D}(S)$ is the set of *probability distributions* on S. For any state $s \in S$, we denote by $A(s)$ the set of actions available from s. We define a *run* of M as a finite or infinite sequence $s_1 a_1 \ldots a_{n-1} s_n \ldots$ 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 $S' \subseteq S$, $A' \subseteq A$, and such that for all $s \in S'$, $A'(s) \neq \emptyset$ and for all $a \in A'(s)$, we have $\mathsf{Supp}(\delta(s, a)) \subseteq S'$, and $\delta'(s, a) = \delta(s, a)$. For all subsets $S' \subseteq S$ with the property that for all $s \in S'$, there exists $a \in A(s)$ with $\mathsf{Supp}(\delta(s, a)) \subseteq S'$, we define the *sub-MDP of* M *induced by* S' as the maximal sub-MDP whose states are S', and denote it by $M|_{S'}$. 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 path. 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 $a \in A(s)$, $\delta(s, a, s) = 1$. We denote by $\mathsf{Abs}(M)$ the set of absorbing states of MDP M.

Histories and 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 $\mathsf{Supp}(\sigma(h)) \in A(s)$. A strategy is *pure* if all histories are mapped to *Dirac distributions*. A strategy σ is *finite-memory* if it can be encoded with a *stochastic Moore machine*, $(\mathcal{M}, \sigma_a, \sigma_u, \alpha)$ where $\mathcal{M}$ is a finite 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 $m \in \mathcal{M}$. A K*-memory strategy* is such that $|\mathcal{M}| = K$. A *memoryless strategy* is such that $|\mathcal{M}| = 1$, and thus only depends on the last state of the history. We define such strategies as functions $s \mapsto \mathcal{D}(A(s))$ for $s \in S$. An MDP M, a finite-memory 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 $\mathbb{P}_{M,s}^\sigma[E]$ the probability of the runs of M_s^σ whose projection to S 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 $h \in (SA)^* S'$, $\mathsf{Supp}(\sigma(h)) \subseteq A'(s)$.

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 a slight 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, $\mathbb{P}_{M,s}^\sigma[\mathsf{Inf} \in \mathsf{MEC}(M)] = 1$ [6]. We call a subset of states *transient* if it is visited finitely many times with probability 1 under any strategy.

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 $\mathsf{Val}_\Phi^\sigma(M, s) = \mathbb{P}_{M,s}^\sigma[\Phi]$ and $\mathsf{Val}_\Phi^*(M, s) = \sup_\sigma \mathbb{P}_{M,s}^\sigma[\Phi]$. 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, $\mathsf{Val}_\Phi^\sigma(M, s) \geq \alpha$. If Φ is achieved with probability 1, we say that it is *achieved almost surely*. Objective Φ is *achieved limit-surely* if for any $\epsilon > 0$, it is achieved 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]. In particular, there exists a strategy ensuring Φ almost-surely when started from any state of $\mathsf{AS}(M, \Phi)$. 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 [17, 5]. The algorithm for parity objectives is obtained by showing that in each end-component the probability of ensuring the objective is either 0 or 1, and then reducing the problem to the reachability of those *winning* end-components. In the next lemma, we recall that the classification of winning end-components does not depend on the exact values of the probabilities, but only on the support of the 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 probabilities 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, for readability, we will study 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 a MEMDP M, state s_0, rationals α_1, α_2, and objective Φ, which is a reachability, safety, or a parity objective, compute a strategy σ, if it exists, such that $\forall i = 1, 2, \mathsf{Val}^{\sigma}_{\Phi}(M_i, s) \geq \alpha_i$.*

We refer to the general problem as *quantitative reachability (resp. safety, parity)*. For an instance 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. Note that in MDPs and MEMDPs, almost-sure safety coincides with *sure* safety (requiring that all runs compatible with a given strategy stay in the safe set of states).

Strategy Complexity We note that 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 componentwise maximal). All counterexamples are given in Fig. 1.

Lemma 3. *For some MEMDPs M and reachability 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).*

The first item is clear from Fig. 1, while the second item follows from the results of the paper. The third item is implies by the next lemma.

Lemma 4. *In the MEMDP M of Fig. 1b, for the reachability objective* $\mathsf{Reach}(T)$*, there exists a Pareto-optimal vector of probabilities achievable by an infinite-memory strategy but not by any finite-memory strategy.*

Proof. Clearly, u is almost surely reached under any strategy. Let us denote $H = (sa + sata)^* u$ the set of histories in M reaching u. Observe also that the probabilities of histories H do not depend on the strategy. Let $P_i(w)$ denote the probability of history $w \in H$ in M_i. We define σ_∞ for any history w with $w \in H$ as a if $P_1(w) \geq P_2(w)$ and as b otherwise.

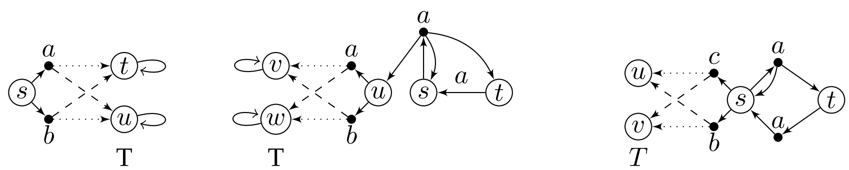
We first show that $\sum_{i=1,2} \mathbb{P}^{\sigma_\infty}_{M_i,s}[\phi] = \sup_\sigma \sum_{i=1,2} \mathbb{P}^{\sigma}_{M_i,s}[\phi]$, where $\phi = \mathsf{Reach}(T)$, which proves that σ_∞ achieves a Pareto-optimal probability vector. In fact, we have for any σ that $\mathbb{P}^{\sigma}_{M_i,s}[\phi] = \sum_{w \in H} \mathbb{P}^{\sigma}_{M_i,s}[\phi \mid w] P_i[w]$. So we get $\mathbb{P}^{\sigma}_{M_1,s}[\phi] = \sum_{w \in H \cap \sigma^{-1}(\{a\})} P_1[w]$ and $\mathbb{P}^{\sigma}_{M_2,s}[\phi] = \sum_{w \in H \cap \sigma^{-1}(\{b\})} P_2[w]$. Since $H \cap \sigma^{-1}(\{a\})$ and $H \cap \sigma^{-1}(\{b\})$ paritions H, we get that $\sum_{i=1,2} \mathbb{P}^{\sigma}_{M_i,s}[\phi] = \sum_{w \in H} P_{f(w)}(w)$, where $f(w) = 1$ if $w \in \sigma^{-1}(\{a\})$ and 2 otherwise. On the other hand, by definition of σ_∞, we have $\sum_{i=1,2} \mathbb{P}^{\sigma_\infty}_{M_i,s} \max(P_1(w), P_2(w))$. Since $Q(w) \leq \max(P_1(w), P_2(w))$ it follows that $\sum_{i=1,2} \mathbb{P}^{\sigma_\infty}_{M_i,s} \geq \sum_{i=1,2} \mathbb{P}^{\sigma}_{M_i,s}$.

Let us now show that no finite-memory strategy achieves $\sup_\sigma \sum_{i=1,2} \mathbb{P}^{\sigma}_{M_i,s}[\phi]$. Consider any m-memory strategy σ for arbitrary $m > 0$. Assume w.l.o.g. that $P_1(sas) > P_2(sas)$. Fix $n = m^3$. Since σ is finite-memory, there exists $0 \leq k_1 < k_2 < m$ such that σ has the same memory element after reading words $(sa)^n(sata)^{k_1}u$ and $(sa)^n(sata)^{k_2}u$. Let us write $w_{n,k} = (sa)^n(sata)^{k_1}u$. We have $\sigma(w_{n,k_1}) = \sigma(w_{n,k_2}) = \alpha \in \{a, b\}$. If $\alpha = b$, then define σ' identically as σ except for $\sigma'(w_{n,k_1}) = a$. We have $P_1(w_{n,k_1}) > P_2(w_{n,k_2})$ so by the above calculations, σ' achieves a higher objective than σ. Assume that $\alpha = a$. In this case, we consider l large enough such that $P_2(w_{n,k_1+l(k_2-k_1)}) > P_1(w_{n,k_1+l(k_2-k_1)})$. This holds for all large enough l since $P_2(sat) > P_1(sat)$. Moreover, on any word $\sigma(w_{n,k_1+l(k_2-k_1)}) = a$ by the above pumping argument. If we define σ' by switching to b at this history, we again improve the objective function, similarly as above. ☐

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

(A) The almost-sure reachability, safety, and parity problems are decidable in polynomial time (Theorems 8 and 32). Finite-memory strategies suffice.

(B) The limit-sure reachability, safety, and parity problems are decidable in polynomial time (Theorem 22 and 39). Moreover, for any $\epsilon > 0$, to achieve probabilities of



(a) (b) (c)

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. 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 b or c 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 (See Lemma 4 in Appendix). 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 estimating the actual environment against which the controller is playing. However, the vector $(1, 1)$ is not achievable, which follows from Section 4.

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 result. We call a MEMDP *acyclic* if the only cycles are self-loops in all environments.

(C) The quantitative reachability and safety problems are NP-hard on acyclic MEMDPs both for arbitrary and memoryless strategies (Theorem 23).

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 represented in unary, the quantitative reachability and safety problems restricted to K-memory strategies can be solved in PSPACE (Theorem 26).

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 39).

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 27).

We will derive approximation algorithms in the following sense.

Definition 5. *The* ϵ-gap problem for reachability *consists, given MEMDP* M*, state* s*, target set* T*, and probabilities* α_1, α_2*, in answering*

– *YES if* $\exists\sigma, \forall i = 1, 2, \mathbb{P}^{\sigma}_{M_i,s}[\mathsf{Reach}(T)] \geq \alpha_i$,
– *NO if* $\forall\sigma, \exists i = 1, 2, \mathbb{P}^{\sigma}_{M_i,s}[\mathsf{Reach}(T)] < \alpha_i - \epsilon$,
– *and arbitrarily otherwise.*

The ϵ-gap problem is an instance of promise problems which guarantee a correct answer in two disjoint sets of inputs, namely positive and negative instances – which do not necessarily cover all inputs, while giving no guarantees in the rest of the input [8, 10].

We give a procedure for the ϵ-gap problem and show its NP-hardness:

(G) There is a procedure for the ϵ-gap problem for quantitative reachability in MEMDPs that runs in double exponential space, and whenever it answers YES, returns a strategy σ such that $\mathbb{P}^{\sigma}_{M,s}[\mathsf{Reach}(T)] \geq \alpha_i - \epsilon$ (Theorem 28).

(H) The ϵ-gap problem is NP-hard (Theorem 29).

Preprocessing Clearly, in a MEMDP, 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, we say that 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 6 (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, *and will be denoted* $R_i(M)$. *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 6 can be made without loss of generality by redirecting any revealing edge to fresh absorbing states. In fact, given an arbitrary MEMDP M, for any objective Φ, we can define M' by replacing any i-revealing edge (s, a, s') in M by two edges $(s, a, \top_i)$ and $(s, a, \perp_i)$ where $\top_i$ (resp. $\perp_i$) is a fresh absorbing winning (resp. losing) state. Here, by winning, we mean that we add $\top_i$ (resp. $\perp_i$) to the set of target (resp. non-target) states for reachability objectives, to the set of safe (resp. unsafe) states for safety objectives, and we assign an even (resp. odd) parity for parity objectives. The probabilities are defined as follows: $\delta'_i(s, a, \top_i) = \delta_i(s, a, s') \cdot \mathsf{Val}^*_{\Phi}(M_i, s')$, and $\delta'_i(s, a, \perp_i) = \delta_i(s, a, s') \cdot (1 - \mathsf{Val}^*_{\Phi}(M_i, s'))$, while the probabilities of other edges are preserved. The interpretation of these values is that at state s, given action a, $\delta_i(s, a, s') \cdot \mathsf{Val}^*_{\Phi}(M_i, s')$ is the probability of going to s', and from thereon winning under the optimal strategy for M_i. The construction is illustrated in Fig. 2.

Note that from any strategy σ' in M' one can derive, by adding one bit of memory, a strategy σ for M such that $\mathbb{P}^{\sigma}_{M_i,s_0}[\Phi] = \mathbb{P}^{\sigma'}_{M'_i,s_0}[\Phi], \forall i = 1, 2$, and $\mathbb{E}^{\sigma}_{M_i,s_0}[\Phi] = \mathbb{E}^{\sigma'}_{M'_i,s_0}[\Phi], \forall i = 1, 2$ respectively for considered objectives. Similarly, any strategy in M can be adapted to M' preserving the probabilities of satisfying a given objective.

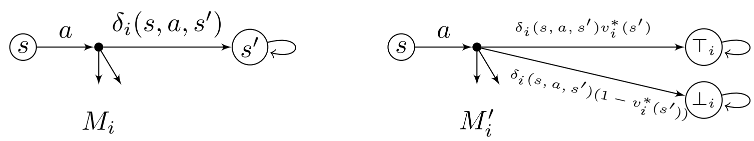


Figure 2: The transformation of any i-revealing edge (s, a, s') so as to put the MEMDP in revealed form, where $v_i^*(s') = \mathsf{Val}_\Phi^*(M_i, s')$, for considered objective Φ.

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

Assumption 7. *For all considered objectives* $\mathsf{Reach}(T)$ *and* $\mathsf{Safe}(T')$*, we assume that* T *and* $S \setminus T'$ *are sets of absorbing states for both environments.*

Under assumptions 6 and 7, 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 $R^\Phi(M) = R_1^\Phi(M) \cup R_2^\Phi(M)$.

Overview We will first concentrate on results on reachability objectives since they contain most of the important ideas. We present algorithms for almost-sure reachability (Section 4), introduce and study *double end-components* (Section 5), then present our algorithms for limit-sure problems (Section 6), and the general quantitative case where we also present NP-hardness results (Section 7). We then summarize our results on safety, and parity objectives (Section 8).

4 Almost-Sure Reachability

We give polynomial-time algorithms for almost-sure reachability in MEMDPs. 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. Formally, for any $s \in S$ and $a \in A(s)$, $\mathsf{Supp}(\delta(s,a)) = \mathsf{Supp}(\delta_1(s,a)) \cup \mathsf{Supp}(\delta_2(s,a))$ and for any $s' \in \mathsf{Supp}(\delta(s,a))$, $\delta(s,a,s') = \frac{1}{|\mathsf{Supp}(\delta(s,a))|}$.

Observe that for any MEMDP M, and subset of states S', the set of states s such that $\mathbb{P}^\sigma_{\cup M,s}[\mathsf{Safe}(S')] = 1$ for some σ induces a sub-MDP in M_1 and M_2. One can therefore define M' the MEMDP induced by this set. Furthermore, any strategy compatible with M' satisfies $\mathsf{Safe}(S')$ surely in each M_i.

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, except for j-revealed states which only need to be winning for M_j. We consider MEMDP M' induced by the states surely satisfying $\mathsf{Safe}(U)$ in both environments. The problem is then reduced to finding strategies in each M_i'. If such strategies we

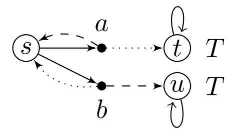


Figure 3: MEMDP M where $\mathsf{Reach}(T)$ can be achieved almost surely. In fact, $\mathsf{AS}(M_i, T) = \{s, t, u\}$ for all $i = 1, 2$, so $M' = M$, and $\mathsf{Val}_{\mathsf{Reach}(T)}(M'_i, s) = 1$ for $i = 1, 2$. The strategy returned by the algorithm consists in choosing, at s, a and b uniformly at random. Notice that there is no pure memoryless strategy achieving the objective almost surely.

obtain our strategy either 1) alternating between two strategies using memory, or 2) randomizing between them. Figure 3 is an example where almost-sure reachability holds. We already saw the example of Fig. 1c where almost-sure reachability does not hold. In fact, in that example M' contains both states $\{s, t\}$ both no winning strategy exists in M'_i for both $i = 1, 2$.

Input: MEMDP M, $\mathsf{Reach}(T)$, $s_0 \in S$
$U := \big(\mathsf{AS}(M_1, \mathsf{Reach}(T)) \cap \mathsf{AS}(M_2, \mathsf{Reach}(T))\big) \cup R^{\mathsf{Reach}(T)}$;
M' := Sub-MEMDP of M induced by states s s.t. $\mathsf{Val}^*_{\mathsf{Safe}(U)}(\cup M, s) = 1$;
if $\forall i = 1, 2, \mathsf{Val}^*_{\mathsf{Reach}(T)}(M'_i, s_0) = 1$ **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 given MEMDP M, starting state s_0 and objective $\mathsf{Reach}(T)$.

Theorem 8. *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 strategy.*

Proof. **(Soundness)** Assume that $\forall i = 1, 2, \mathsf{Val}^*_{\mathsf{Reach}(T)}(M'_i, s) = 1$, and consider pure memoryless strategies σ_i achieving $\mathsf{Reach}(T)$ almost surely in each M'_i from any state of U, and let $\sigma = \frac{1}{2}\sigma_1 + \frac{1}{2}\sigma_2$. We have $\mathbb{P}^{\sigma}_{M_i,s}[\mathsf{Safe}(U)] = 1$ for any i since each σ_i is compatible with M'_i. Moreover, for each i, and from any state s' of M_i reachable under σ, target set T is reached with positive probability in $|S|$ steps under strategy σ_i. In fact we have, for such a state s', $s' \in U \setminus R_{3-i}$. Since the probability of σ being identical to σ_i for $|S|$ steps is positive, T is reached almost surely in M_i under σ from s.

This construction gives a memoryless strategy. One can obtain a pure finite-memory strategy by alternating between σ_1 and σ_2 every $|S|$ steps.

(Completeness) Conversely, assume that there exists a strategy σ almost surely achieving $\mathsf{Reach}(T)$ from s. Towards a contradiction, assume that $\mathsf{Val}^{\sigma}_{\mathsf{Safe}(U)}(M_i, s) <$

1. This means some state $t \notin U$ is reached with positive probability under σ. Recall that all target states are absorbing in M by Assumption 7. If $t \in R_i \setminus R_i^{\mathsf{Reach}(T)}$ this contradicts that σ almost surely achieves the objectives, and similarly if $t \notin \mathsf{AS}(M_i, T)$, since a target state could not have been reached before arriving to t. Last, if $t \notin \mathsf{AS}(M_{3-i}, T)$ and t is not revealed, then this state is also reachable with positive probability in M_{3-i} under σ, which is again a contradiction. Therefore $\mathsf{Val}^{\sigma}_{\mathsf{Safe}(U)}(M_i, s) = 1$ for all $i = 1, 2$, which means that s is a state of M' and σ is compatible with M'. Last, we do have $\mathsf{Val}^{*}_{\mathsf{Reach}(T)}(M'_i, s_0) = 1$ since σ is a witnessing strategy. Therefore, the algorithm answers positively on this instance. ☐

5 Double end-components

End-components play an important role in the analysis of MDPs [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. We show that one can *learn* inside double end-components, and use these observations to study limit-sure objectives.

Formally, given a 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 exists $(s, a) \in S' \times A'$ such that $\delta_1(s, a) \neq \delta_2(s, a)$. The union of two DECs with a common state is a DEC; we consider maximal DECs (MDEC). MDECs can be computed in polynomial time by first eliminating from M all actions with different supports, and then computing the MECs in the remaining MDPs. A DEC is *trivial* if it is an absorbing state.

Under Assumption 7, for reachability objectives, a DEC is *winning* if it is an absorbing state winning for the objective. A DEC D is *winning* for a parity objective Φ, if there exists a strategy compatible with D satisfying Φ almost surely; Lemma 1 shows that a common strategy exists for both environments.

We first solve the problems of interest in distinguishing DECs up to any error bound ϵ. The idea is that in a distinguishing DEC, one can learn the environment by sampling the distribution of distinguishing actions.

Lemma 9. *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 $\mathbb{P}^{\sigma}_{M_i,s}[\Phi] \geq (1 - \epsilon)\mathsf{Val}^{*}_{\Phi}(M_i, s), \forall i = 1, 2$.*

Proof. Fix $(s, a) \in S' \times A'$ such that $\delta_1(s, a) \neq \delta_2(s, a)$. The strategy runs in two rounds. In the first round, the goal is to sample the distribution of the edge (s, a). For this, it suffices to execute a strategy that chooses each available action compatible with D uniformly at random, and upon arrival to state s, to choose action a, and store the number of times the next state is s'. After K visits to s, we make a guess about the current MDP depending on the sampled value. The second round of the strategy is the memoryless optimal strategy in one of the M_i. When K is chosen sufficiently large, we obtain the desired result.

Let us denote $d_i = \delta_i(s, a, s')$ for some s' satisfying $d_1 \neq d_2$, and assume w.l.o.g. that $d_1 < d_2$. For any $\epsilon > 0$, let $K = 2\frac{\log(1/\epsilon)}{(d_2-d_1)^2}$, and let f be a memoryless strategy which chooses uniformly at random all actions except action a is picked at s deterministically. Under f, each state is visited infinitely often almost surely. We define f_K by augmenting f with memory as follows. Informally, f_K has two counters: $c_{s,a}$ counting the number of visits at s, and $c_{s,a,s'}$ counting the occurence of edge (s, a, s'). Hence, at each visit at s, we have a Bernouilli trial with mean $\delta_i(s, a, s')$ (for each i), and $c_{s,a,s'}$ is the number of successful trials. It is clear that the ratio $c_{s,a,s'}/c_{s,a}$ should go to $\delta_i(s, a, s')$ inside each M_i. We execute this strategy until $c_{s,a} = K$, which happens almost surely. We complete the description of strategy f_K by extending it, once $c_{s,a,s'} = K$ is reached, with the optimal memoryless strategy opt_1 for M_1 if $\frac{c_{s,a,s'}}{c_{s,a}} \leq \frac{d_1+d_2}{2}$, and opt_2, the one for M_2 otherwise.

By Hoeffding's inequality, we have

$$\mathbb{P}^{f_K}_{M_1,s}[\frac{c_{s,a,s'}}{c_{s,a}} \geq d_1 + \frac{d_2-d_1}{2} \mid c_{s,a} = K] \leq e^{-2K\frac{d_2-d_1}{2}^2} \leq \epsilon.$$

and

$$\mathbb{P}^{f_K}_{M_2,s}[\frac{c_{s,a,s'}}{c_{s,a}} \leq d_2 - \frac{d_2-d_1}{2} \mid c_{s,a} = K] \leq e^{-2K\frac{d_2-d_1}{2}^2} \leq \epsilon.$$

We now compute the values under strategy f_K, distinguishing whether the sampled frequency stays within the given radius or not. In the first case, the objective is satisfied with probability $\mathsf{Val}^*_\Phi(D)$, and in the second case, with probability at least 0. It follows that $\mathbb{P}^{f_K}_{M_i,s}[\Phi] \geq (1-\epsilon)\mathsf{Val}^*_\Phi(D)$. Note that the memory requirement is K^2, since we store the pairs $(c_{s,a}, c_{s,a,s'})$. □

Remark 10. *The algorithm can be improved in practice as follows. Let S' denote the set of states of the end-component which have distinguishing actions. For any state $s \in S'$, fix a distinguishing action a_s. For any s' such that $\delta_1(s, a_s, s') \neq \delta_2(s, a_s, s')$, write $K_{s,a_s,s'}$ the above constant computed for this edge. We apply the following strategy: at any state $s \in S'$ play a_s, and sample the distribution. At any state $s \notin S'$, pick an action uniformly at random. Now, we run this strategy until we collected $K_{s,a_s,s'}$ samples for some action a_s. Note that if S' is a singleton, this does not improve the lemma's proof.*

What expected time can we guarantee until the environment is guessed with prob. $1-\epsilon$? Let $T(s', s)$ denote the expected time to reach state s from s' under the uniform strategy [2], and let $T(s) = \max_{s'} T(s', s)$. If s denotes a state with a distinguishing action, such that $\eta = |\delta_1(s, a, s') - \delta_2(s, a, s')|$, then the above algorithm switches to a pure optimal strategy in expected $O(T(s)\frac{\log(1/\epsilon)}{\eta^2})$ time.

We now consider general MEMDPs, and define a transformation by contracting DECs. The transformation preserves, up to any desired ϵ, the probabilities of objectives, thanks to Lemma 9.

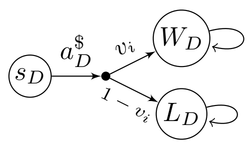
[2]Note that since we do not know the exact distributions, we cannot minimize the expected time using an optimal strategy here.

Given a DEC $D = (S', A')$, a *frontier state* s of D is such that there exist $a \in A(s) \setminus A'(s)$, $i \in \{1, 2\}$, 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.

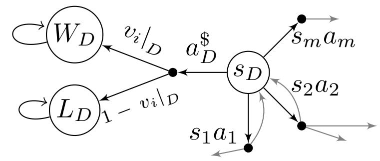
Definition 11. *Given a MEMDP $M = (S, A, \delta_1, \delta_2)$, and reachability or safety objective Φ, we define $\hat{M} = (\hat{S}, \hat{A}, \hat{\delta_1}, \hat{\delta_2})$ as follows. a) Any distinguishing MDEC D is contracted as in Fig. 4a where in M_i, action a leads to new states W_D with probability $v_i = \mathsf{Val}^*_\Phi(M_i, D)$, and to L_D with probability $1 - v_i$. b) Any non-distinguishing MDEC $D = (S', A')$ is replaced with the module in Fig. 4b. 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'.*

We 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 denote by $\hat{\mathcal{A}} : S \to \hat{S}$ the mapping from the states of S to that of $\hat{S}$ defined by the above transformation, mapping any state s of a DEC D is to s_D, and any other state to itself. We will also denote $\hat{s} = \hat{\mathcal{A}}(s)$.



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



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

The intuition is that when the play enters a distinguishing DEC D, by applying Lemma 9, we can arbitrarily approximate probabilities $v_i = \mathsf{Val}^*_\Phi(M_i, D)$. From a state s in a non-distinguishing component D in M, the play either stays forever inside and obtain the value $\mathsf{Val}^*_\Phi|_D\,(M_1, s) = \mathsf{Val}^*_\Phi|_D\,(M_2, s)$ (as it is non-distinguishing), or it eventually leaves D. The first case is modeled by the action $a^\$_D$, and the second case by the remaining actions leading to frontier states. Note that there is a strategy under which, from any state of D, in M_1 and M_2, all states and actions of D are visited infinitely often (by considering a memoryless strategy choosing all actions uniformly at random – see *e.g.* [17]). We will use this construction for reachability and safety objectives; while a specialized construction based on $\hat{M}$ will be defined for parity objectives.

The point in defining $\hat{M}$ is to eliminate all non-trivial DECs:

Lemma 12. *Let D be a maximal end-component of $\hat{M}_i$. Then either D is a trivial DEC, or D is transient in $\hat{M}_{3-i}$.*

Proof. Assume that D is an end-component of $\hat{M}_{3-i}$. Then D is a double end-component by definition. If D is a self-loop, then it is an absorbing state and we are done. Otherwise, D must contain some state s_E of $\hat{M}$ created by contracting MDEC E since otherwise D would have been contracted itself by definition of $\hat{M}$. But then $D \cup E$ is a DEC larger than D, which is a contradiction. Thus, D cannot be an end-component of $\hat{M}_{3-i}$ unless it is one absorbing state.

Assuming D is not an end-component of $\hat{M}_{3-i}$, either D is not strongly connected, or it is not δ_{3-i}-closed. Observe first that D does not contain i-revealing edges in $\hat{M}_i$ since otherwise, by construction of $\hat{M}_i$, it contain an absorbing state and not be strongly connected in $\hat{M}_i$. We show that D must also be strongly connected in $\hat{M}_{3-i}$. In fact, assume otherwise and consider two states s and t such that t is not reachable from s in $\hat{M}_{3-i}\Big|_D$. Along the run from s to t, $\hat{M}_i$ must have an edge that is absent from $\hat{M}_{3-i}$, which is an i-revealing edge; contradiction. Therefore, D is strongly connected and not δ_{3-i}-closed in $\hat{M}_{3-i}$.

We now show that under any strategy in $\hat{M}_{3-i}$, the play eventually leaves D almost surely. It suffices to show that $\hat{M}_{3-i}$ has no end-component inside D. Let $D' \subseteq D$ be such an end-component. Then D' does not contain $3-i$-revealing edges; in fact, we know that D is strongly connected, and a $3-i$-revealing edge means an absorbing state inside D. Note that D' does not contain $3-i$-revealing state-actions neither since these would lead outside D, and D' would not be δ_{3-i}-closed. This means that the sub-MDP D' has the same support in both $\hat{M}_j$, hence it is also an end-component of $\hat{M}_i$, hence D' is a double end-component. But this is only possible, by construction of $\hat{M}$, if $D = D'$ is an absorbing state. ☐

The following lemma refines the above one.

Lemma 13. *For any M, and $\epsilon > 0$, define $K = n\lceil \frac{\log(\epsilon)}{\log(1-p^n)} \rceil$, where p is the smallest nonzero probability of M, and n the number of states. for any end-component D of $\hat{M}_i$ that is not a DEC, and any history $h \in \mathcal{H}(\hat{M})$ which contains a factor of length K compatible with D, $\mathbb{P}^{\tau}_{\hat{M}_{3-i},s}[h] \leq \epsilon$ for any strategy τ and state s.*

Proof. We know that D does not contain an end-component in $\hat{M}_{3-i}$. If p denotes the smallest nonzero probability in $\hat{M}$, then from any state $s \in D$, the probability of leaving D after n steps is at least p^n under any strategy. So in K steps, the probability of leaving D is at least $\sum_{i=0}^{K/n}(1-p^n)^i p^n = \frac{1-(1-p^n)^{K/n+1}}{p^n} p^n = 1-(1-p^n)^{k+1}$. which is at least $1-\epsilon$. ☐

In order to prove the "equivalence" of M and $\hat{M}$ for objectives of interest, we define a correspondance between histories of M and $\hat{M}$ which is, roughly, the projection defined by our transformation. We distinguish the set $\mathcal{T}(\hat{M}) = \{s_D \mid D \text{ distinguishing}\}$. For any history $h = s_1a_1s_2a_2\ldots s_n \in \mathcal{H}(M)$, let us define $\mathsf{red}(s_1a_1s_2a_2\ldots s_n) \in \mathcal{H}(\hat{M})$ by applying the following transformations until a fixpoint is reached:

1. If h contains a state of $\hat{\mathcal{A}}^{-1}(\mathcal{T}(\hat{M}))$, then if i denotes the least index with $s_i \in \hat{\mathcal{A}}^{-1}(\mathcal{T}(\hat{M}))$, we remove the suffix $a_is_{i+1}\ldots s_n$.
2. For any non-distinguishing MDEC D, let $s_ia_i\ldots s_{i+k}$ be a maximal factor made of the states of D. We remove from this factor all non frontier actions and states that precede. We project all states to s_D, and any action a_{α_j} from state s_{α_j} to

action $(s_{\alpha_j} a_{\alpha_j})$. We obtain a run of the form $s_D(s_{\alpha_1} a_{\alpha_1})s_D \dots s_D(s_{\alpha_m} a_{\alpha_m})$ where each s_{α_i} is a frontier state, and a_{α_i} a frontier action from s_{α_i}.

Let $\mathcal{H}_{\mathcal{T}}(\hat{M})$ denote the histories of $\hat{M}$ which does not contain $\mathcal{T}(\hat{M})$ except possibly on the last state. The following lemma establishes the relation between M and $\hat{M}$.

Lemma 14. *For any MEMDP M, state s, strategy σ, there exists a strategy $\hat{\sigma}$ such that for any history $h \in \mathcal{H}_{\mathcal{T}}(\hat{M})$, and any non-distinguishable MDEC D, we have $\mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(h) = \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h)], \mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(ha) = \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(ha)]$, and $\mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}[ha^{\$}_D] = \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(hD^{\omega})]$.*

Proof. Let us restate the equalities we are going to prove.

$$\begin{aligned}
&\mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(h) = \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h)],\\
&\mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(ha) = \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(ha)],\\
&\mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}[ha^{\$}_D] = \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(hD^{\omega})].
\end{aligned} \tag{1}$$

Given σ, we define $\hat{\sigma}$ as follows. For any history ending in $\mathcal{H}_{\mathcal{T}}(\hat{M})$, $\hat{\sigma}$ is defined trivially. For any $a \in \hat{A}(h_i) \setminus \{a^{\$}_D\}_D$, define

$$\hat{\sigma}(a \mid h_1 \dots h_i) = \mathbb{P}^{\sigma}_{M_j,s}(\mathsf{red}^{-1}(h_1 \dots h_i a) \mid \mathsf{red}^{-1}(h_1 \dots h_i)),$$

for an arbitrary j. These quantities do not depend on j. In fact, $\mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(ha) \mid \mathsf{red}^{-1}(h)] = \sum_{\pi \in \mathsf{red}^{-1}(h)} \mathbb{P}^{\sigma}_{M_j,s}[\pi a \mid \pi]\mathbb{P}^{\sigma}_{M_j,s}[\pi \mid h] = \sum_{\pi \in \mathsf{red}^{-1}(h)} \sigma(a \mid \pi)\mathbb{P}^{\sigma}_{M_j,s}[\pi \mid h]$, and the latter factor does not depend on j; since $\mathsf{red}^{-1}(h)$ determines all outcomes of the actions whose distributions differ in both M_i, and the distributions are identical in the remaining non-distinguishing double end components.

For any $h_i = s_D$, where D is a non-distinguishing component, we let

$$\hat{\sigma}(a^{\$}_D \mid h_1 \dots h_i) = \mathbb{P}^{\sigma}_{M_j,s}(\mathsf{red}^{-1}(h_1 \dots h_i D^{\omega}) \mid \mathsf{red}^{-1}(h_1 \dots h_i)).$$

We check that $\hat{\sigma}$ defines a probability distribution on available actions at any given history. For any state $h_i \neq s_D$, probabilities $\hat{\sigma}(a \mid h_1 \dots h_i)$ clearly sum to 1 for $a \in A(h_i)$. If $h_i = s_D$ for some non-distinguishing losing D, any run that extends $h_1 \dots h_i$ either stays forever in D, or takes one of the frontier actions for the first time. By definition, the former is the probability of $\hat{\sigma}$ of choosing $a^{\$}_D$, and the latter that of choosing each frontier action.

We will prove (1) by induction on $i \geq 1$.

For $i = 1$, we have $\mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(h_1) = \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h_1)]$, which is 1 if $\hat{s} = h_1$ and 0 otherwise. Furthermore, $\mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(h_1 a_1) = \mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(h_1)\hat{\sigma}(a_1 \mid h_1) = \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h_1)] \cdot \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h_1 a_1) \mid \mathsf{red}^{-1}(h_1)] = \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h_1 a_1)]$.

For $i > 1$, we have

$$\begin{aligned}
\mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(h_1 \dots h_i) &= \mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}[h_1 \dots h_{i-1} a_{i-1}]\hat{\delta}_j(h_{i-1}, a_{i-1}, h_i)\\
&= \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h_1 \dots h_{i-1} a_{i-1})]\hat{\delta}_j(h_{i-1}, a_{i-1}, h_i)\\
&= \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h_1 \dots h_{i-1} a_{i-1} h_i)].
\end{aligned}$$

The second line follows by induction, and the third line by definition (explain). We have

$$\begin{aligned}\mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(h_1 \dots h_i a_i) &= \mathbb{P}^{\hat{\sigma}}_{\hat{M}_j,\hat{s}}(h_1 \dots h_i)\hat{\sigma}(a_i \mid h_1 \dots h_i) \\ &= \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h_1 \dots h_i)] \\ &\quad \cdot \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(h_1 \dots h_i a_i) \mid \mathsf{red}^{-1}(h_1 \dots h_i)]\end{aligned}$$

The third equality is proved similarly. □

The equivalence between M and $\hat{M}$ for reachability and safety objectives is obtained as the following corollary. Note that the value vectors are preserved although vectors achieved in $\hat{M}$ may not be achievable in M.

Corollary 15. *For any MEMDP M, and Φ a reachability or safety objective,* $\mathsf{Val}^*_{\Phi}(M, s) = \mathsf{Val}^*_{\hat{\Phi}}(\hat{M}, \hat{s})$.

By Definition 11, and the previous corollary, we assume, in the next section, that the MEMDPs we consider have only trivial DECs.

Assumption 16. *All MEMDPs are assumed to have only trivial DECs.*

6 Limit-Sure Reachability

In this section, we give a polynomial-time algorithm for limit-sure reachability in MEMDPs. For any MEMDP M, and reachability objective Φ, we define the set of limit-sure winning states $W(M, \Phi)$ as follows. We have $s \in W(M, \Phi)$ if either $s \in R^{\Phi}$, or there exists a family of strategies witnessing limit-sure satisfaction, that is, for any $\epsilon > 0$, there a strategy σ_ϵ such that $\mathbb{P}^{\sigma_\epsilon}_{M_i,s}[\Phi] \geq 1 - \epsilon$ for $i = 1, 2$.

The following lemma states an important property of the set $W(M, \Phi)$ for reachability objectives but also safety objectives.

Lemma 17. *On any MEMDP M, and a reachability or safety objective Φ, there exists a memoryless strategy σ_W under which from any $s \in W(M, \Phi)$, each M_i stays surely inside $W(M, \Phi)$.*

of Lemma 17. In this proof only, we separate control and probabilistic states for convenience. Given a state s, and action $a \in A(s)$, we denote by sa the intermediate probabilistic state reached by chosing action a. We denote $W = W(M, \Phi)$.

We show that all successors of probabilistic states $sa \in W$ are in W. In fact, assume that there exists $s' \notin W$ such that $\delta_i(s, a, s') \neq 0$ for some i. This means that $s' \notin R^{\Phi}$ and there is no family of strategies witnessing limit-sure winning from s'. If $s' \in R \setminus R^{\Phi}$, then there exists $\epsilon_0 > 0$ such that for any strategy σ, $\mathbb{P}^{\sigma}_{M_i,s'}[\Phi] \leq 1 - \epsilon_0$, therefore $\mathbb{P}^{\sigma}_{M_i,s}[\Phi] \leq 1 - \delta_i(s, a, s') + \delta_i(s, a, s')(1 - \epsilon_0) \leq 1 - \epsilon_0 \delta_i(s, a, s')$ contradicting that $s \in W$. Note that we cannot have $s' \in R_{3-i}$ since $\delta_i(s, a, s') \neq 0$. Now, if s' is unrevealed then $\delta_j(s, a, s') \neq 0$ for both $j = 1, 2$. By assumption that $s' \notin W$, there exists $\epsilon_0 > 0$ such that for any strategy σ, $\mathsf{Val}^*_{\Phi}(M_j, s') \leq 1 - \epsilon_0$ for some j. Then, for any σ, for some j, $\mathbb{P}^{\sigma}_{M_j,s}[\Phi] \leq 1 - \delta_j(s, a, s')\epsilon_0$ contradicting $s \in W$.

We now prove that for any control state $s \in W$, there exists an action a such that $\mathsf{Supp}(s, a) \subseteq W$, by induction on the length $k > 0$ of the history. At the same time, we define the strategy σ_W by setting $\sigma_W(s) = a$.

The case $k = 1$ is trivial since $s \in W$. For probabilistic states sa, the property follows from the above paragraph. Assume $k \geq 2$. If there exists $a \in A(s)$ such that $sa \in W$, then by induction hypothesis, for all $\epsilon > 0$, there exists a strategy σ' that witnesses $1 - \epsilon$-satisfaction from the (probabilistic) state sa and stays in W states for $k - 1$ steps. We let $\sigma_W(s) = a$.

We now prove that there must exist such an action a. To get a contradiction, assume that for all actions $a \in A(s)$, $sa \notin W$. This means that for all $a \in A(s)$, there exists $\epsilon_a > 0$ such that $\mathsf{Val}^*_\Phi(M_j, sa) \leq 1 - \epsilon_a$ for some j. Let σ be a strategy witnessing $1 - \epsilon$-satisfaction for $\epsilon < \frac{1}{|A(s)|} \min_{a \in A(s)} \epsilon_a$. There exists $a \in A(s)$ such that from s, σ assigns a probability of at least $\frac{1}{|A(s)|}$ to a. Let j such that $\mathsf{Val}^*_\Phi(M_j, sa) \leq 1 - \epsilon_a$. We have $\mathbb{P}^\sigma_{M_j,s}[\Phi] \leq \frac{1}{|A(s)|}(1 - \epsilon_a) + 1 - \frac{1}{|A(s)|} < 1 - \epsilon$, contradiction. $\square$

In the rest of the paper, σ_W will denote the pure memoryless strategy of Lemma 17. Note that we do not require the *computability* of σ_W at this point.

In the rest of this section, we assume, by Assumption 16, that the considered MEMDPs have only trivial DECs.

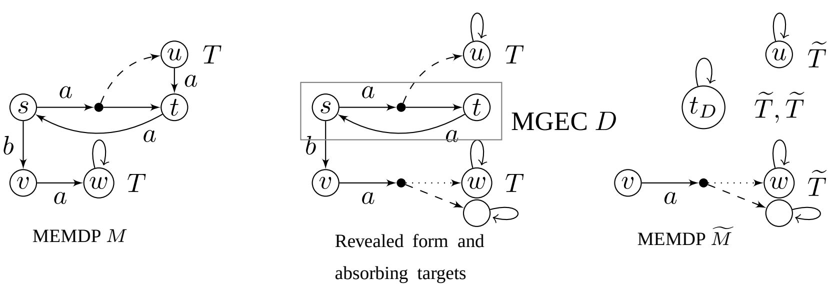


Figure 4: On the left, an MEMDP with objective $\mathsf{Reach}(T)$, which is *not* in revealed form; an equivalent instance M in revealed form is shown in the middle. Note that M has only trivial DECs. States $\{s, t\}$ induce a good end-component D in M_2; in fact, the strategy choosing action a at 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 becomes a target state. Because $\widetilde{\mathcal{A}}(s) = t_D$, objective Φ is achieved limit-surely from s.

Let us explain the idea behind the limit-sure reachability algorithm on the MEMDP M 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 chooses a at states s and t almost surely reaches u in M_2. On order to achieve the objective with probability close to 1, one can run strategy σ for a large number of steps, and if the objective is still not achieved, switch to the optimal strategy for M_1, which consists in choosing b from s. It can be shown that such a strategy achieves the objective at probabilites $(1-\epsilon, 1-\epsilon)$, for any desired $\epsilon > 0$, from any state of such end-components. Our algorithm consists in identifying these end-components and contracting them as winning absorbing states.

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

Lemma 18. *Let M be a MEMDP with only trivial DECs, and a reachability objective Φ. For any $i = 1, 2$, consider the set*

$U_i = \{s \mid \exists D \in \mathsf{MEC}(M_i), s \in D, \mathsf{Val}^*_{\mathsf{Safe}(D \cup R^{\Phi}_{3-i})}(\cup M|_{D \cup R_{3-i}}, s) = 1\}.$

Let M'_i denote the sub-MDP of M_i induced by U_i. Then the MGECs of M_i are the union of the MECS of M'_i, and the trivial MECs of M'_i surely satisfying Φ.

Proof. To see that the sub-MDP M'_i is well-defined, notice that for each D, the states satisfying the safety condition induces a sub-MDP, and that these sub-MDPs are disjoint for each D.

Let us show that non-trivial MECs of M'_i and trivial-and-winning MECs of M'_i are MGECs of M_i. Note that we distinguish here the case of trivial MECs since our definition U_i could yield trivial MECs that are not winning. It is clear that trivial MECs of M'_i satisfying Φ are maximal good end-components. Consider a non-trivial MEC G of M'_i. Let τ be the uniform strategy inside G in M'_i. Clearly, τ stays inside G in M_i. In M_{3-i}, we know that strategy τ leaves G almost surely by Lemma 12. But by Assumption 6, and by the fact that τ is compatible with U, τ also ensures $\mathsf{Safe}(D \cup R^{\Phi}_{3-i})$ surely, so R^{Φ}_{3-i} must be reached almost surely in M_{3-i}. Therefore, G is a good end-component. We will show its maximality at the end of this proof.

Conversely, we show that MGECs of M_i are MECs of M'_is. Any MGEC G of M_i is in particular a MEC of M_i, so it is included in some $D \in \mathsf{MEC}(M_i)$. Let τ be the uniform strategy in G. Clearly, we have $\mathbb{P}^{\tau}_{M_i,s}[\mathsf{Safe}(D)] = 1$ for any $s \in D$, and $\mathbb{P}^{\tau}_{M_{3-i},s}[\mathsf{Safe}(D \cup R^{\Phi}_{3-i})] = 1$. In fact, because strategy τ is compatible with D in M_i, and by Assumption 6, any action of D which leaves D in M_{3-i} ends in R_{3-i}. Furthermore, because τ is almost surely winning for M_{3-i} from D, we have that $\mathsf{Safe}(D \cup R^{\Phi}_{3-i})$ holds surely in M_{3-i} under τ. It follows that G is included in M'_i. Moreover, G is by definition an end-component in M'_i. To show that G is maximal, assume that there exists $G \subsetneq G' \subseteq M'_i$ where G' is a MEC in M_i. By the first case, G' is a good end-component which contradicts the maximality of G as a good end-component. Therefore, G is indeed a MEC of $M_i|_U$.

To finish the proof, we show that a non-trivial MEC G of M'_i is a *maximal* good end-component. Towards a contradiction, assume that there exists $G \subsetneq G'$ a MGEC of M_i. By the second case above, G' is then a MEC of M'_i which contradicts the maximality of G as an end-component of M'_i. ☐

Definition 19 (Transformation $\widetilde{M}$)**.** *Given any MEMDP M with only trivial DECs, and reachability objective Φ. we define $\widetilde{M} = (\widetilde{S}, \widetilde{A}, \widetilde{\delta_1}, \widetilde{\delta_2})$ by applying the following transformation to M. 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.

Let us denote by $\widetilde{\mathcal{A}}(\cdot)$ the mapping from the states M to those of $\widetilde{M}$.

The following lemmas establish the equivalence between limit-sure objectives in M and corresponding almost-sure objectives in $\widetilde{M}$. The algorithm for limit-sure objectives is then obtained by using the algorithm of Section 4. Note that only the first lemma is constructive, but it is the one that we need to compute strategies for M.

Lemma 20. *For any MEMDP M with only trivial DECs, and reachability objective Φ, if $\widetilde{\Phi}$ can be achieved almost surely in $\widetilde{M}$, then Φ can be achieved limit surely in M. Moreover, given an almost sure winning strategy for $\widetilde{M}$, for any $\epsilon > 0$, a strategy with memory $O(\frac{\log(\epsilon)}{\log(1-p)})$ for M, where p is the smallest nonzero probability, achieving probabilities $1 - \epsilon$ can be computed.*

Proof. Let σ be a strategy achieving each $\widetilde{\Phi}$ almost surely in $\widetilde{M}$. For any $\epsilon > 0$, we derive a strategy for M achieving Φ with probability $1 - \epsilon$ for each M_i. For this, we define σ_ϵ for M by modifing σ as follows. Remember that all target states are absorbing by Assumption 7. Fix any $\epsilon \in [0, 1]$, and let p be the smallest nonzero probability in M. Define $K \geq \frac{\log(\epsilon)}{\log(1-p)}$. Upon arrival to any state of a MGEC D of M_j, if D is a trivial DEC, then we extend the strategy trivially. Otherwise, we switch to a strategy τ compatible with D in M_j picking all actions in D uniformly at random. Note that under this strategy, actions B that leave D in M_{3-j} with positive probability are seen infinitely often (since D is not a DEC). These actions lead to $3 - j$-revealed states in M_{3-j} from which the strategy is extended trivially. Whenever actions in B are seen K times, if the play is still in D then we switch to the optimal strategy for M_j. Notice that the probability of staying inside D under τ in M_j is 1, while the probability of leaving D in M_{3-j} under strategy is at least $1 - \epsilon$ by the choice of K.

Assume $\widetilde{\Phi} = \mathsf{Reach}(T \cup \{t_D\})$. Because $\widetilde{\Phi}$ is ensured almost surely in $\widetilde{M}_i$, in M_i under σ, almost surely we either reach T or switch to τ. The claim follows since when we switch to τ, T is reached with probability at least $1 - \epsilon$. ☐

Lemma 21. *Let M be any MEMDP with only trivial DECs, and Φ reachability objectives. Let σ_W denote the strategy of Lemma 17 for M, and $\widetilde{\sigma}_W$ obtained from σ_W by extending it trivially on states t_D. For any $s \in W(M, \Phi)$, $\mathsf{Val}_{\widetilde{\Phi}}^{\widetilde{\sigma}_W}(\widetilde{M}, \widetilde{s}) = 1$.*

of Lemma 21. For strategy $\widetilde{\sigma}_W$ and starting state $\widetilde{s}$, let D be any MEC of $\widetilde{M}_i$ in which the play stays forever with positive probability. We have $D \subseteq \widetilde{\mathcal{A}}(W(M, \Phi)) \cup \{t_D\}_D$ since σ_W does not leave the set $W(M, \Phi)$ in M. If D is a DEC, then it is trivial and satisfies the objective. If D is not a DEC, then it is transient in $\widetilde{M}_{3-i}$. But because σ_W does not leave the set $W(M, \Phi)$, all revealed states reached under σ_W from D in $\widetilde{M}_{3-i}$ are in R_{3-i}^{Φ}, therefore winning. It follows that D is a good end-component, contradiction since all such components were reduced in $\widetilde{M}$. Therefore, any MEC D of $\widetilde{M}_i$ in which the play stays forever is a DEC satisfying $\widetilde{\phi}_i$. The lemma follows. ☐

The algorithm consists in constructing $\widetilde{M}$ and solving almost-sure reachability for $\widetilde{\Phi}$:

Theorem 22. *The limit-sure reachability problem is decidable in polynomial-time.*

7 Quantitative Reachability

We are now interested in the general quantitative reachability problem for MEMDPs. We first show that the problem is NP-hard, so it is unlikely to have a polynomial-time algorithm, and techniques based on linear programming cannot be applied. We will then derive an approximation algorithm.

7.1 Hardness

The main hardness result is the following.

Theorem 23. *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 $i = 1, 2$.*

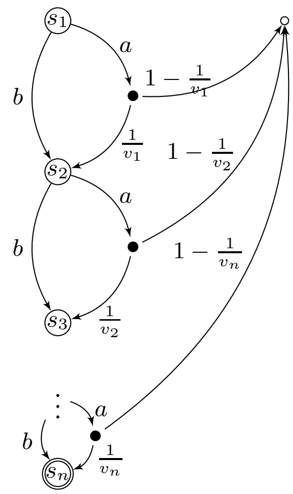
We prove this theorem by reduction from the product-partition problem [15]. The same reduction will also yield the hardness of the ϵ-gap problem, which we show at the end of this subsection.

The following Product-Partition problem is NP-hard in the strong sense. Given positive integers $v_1, \ldots, v_n$, decide whether there exists a subset $I \subseteq \{1, \ldots, n\}$ such that $\prod_{i \in I} v_i = \prod_{i \notin I} v_i$. It is easy to see that the problem is equivalent if the target value $\sqrt{v_1 \ldots v_n}$ is given as part of input. In fact, if l is the maximum number of bits required to represent any v_i, then $V = v_1 \ldots v_n$ can be computed in time $n^2 l^2$. Further, one can check if V is a perfect square and (if it is) compute the square root in time $O(\log(V)) = O(nl)$ by binary search.

We reduce this problem to quantitative reachability in MEMDPs. We fix an instance of the problem, and construct the following MEMDP. The figure depicts the MDP M_1, while M_2 is obtained by inversing the roles of a and b. We let s_n be the target state. Let us denote $W = 1/V$. We will prove that M has a strategy achieving the probabilities $(\sqrt{W}, \sqrt{W})$ for reaching s_n if, and only if the Product-Partition problem has a solution. Notice that the reduction is polynomial since all probabilities can be encoded in polynomial time.

Observe that each pure strategy σ corresponds a set $S_\sigma = \{i \mid \sigma(s_i, b) = 1\}$. We have that $\mathbb{P}^{\sigma}_{M_2,s_1}[\mathsf{Reach}(T)] = \prod_{i \in S_\sigma} \frac{1}{v_i}$, and $\mathbb{P}^{\sigma}_{M_1,s_1}[\mathsf{Reach}(T)] = \prod_{i \notin S_\sigma} \frac{1}{v_i}$ Therefore, a pure strategy with values $(\sqrt{W}, \sqrt{W})$ yields a solution to the Product-Partition problem, and conversely. To establish the reduction, we need to show that if some arbitrary strategy achieves the probability vector $(\sqrt{W}, \sqrt{W})$ in M, then there is a pure strategy achieving the same vector.

To ease reading, for any strategy σ, let us denote $p_i^\sigma = \mathbb{P}^{\sigma}_{M_i,s_1}[\mathsf{Reach}(T)]$. Note that strategies do not need memory on this MEMDP: any state s_i uniquely determines the history. We thus concentrate on randomized memoryless strategies. Let Σ^D denote the set of deterministic strategies.



Claim: For any stationary strategy σ, there exists $(\lambda_\pi)_{\pi\in\Sigma^D}$ with $0 \leq \lambda_\pi \leq 1$ and $\sum_{\pi\in\Sigma^D} \lambda_\pi = 1$ such that $p_i^\sigma = \sum_{\pi\in\Sigma^D} \lambda_\pi p_i^\pi$ for all $i = 1, 2$.

The claim can be proved as follows. There is a finite number of deterministic strategies, each corresponding to a word in $(a + b)^n$. By summing over all runs leading to s_n, we write $p_j^\sigma = \sum_{\pi\in(a+b)^n} \prod_{i=1}^n \sigma(s_i, \pi_i)\delta_j(s_i, \pi_i, s_{i+1})$ so we let $\lambda_\pi = \prod_{i=1}^n \sigma(s_i, \pi_i)$, and observe that $\prod_{i=1}^n \delta_j(s_i, \pi_i, s_{i+1}) = p_j^\pi$. $\square$

The following lemma is the last step of the reduction: if there is a strategy achieving a probability vector close to $(\sqrt{W}, \sqrt{W})$ by some ϵ, then there is a pure strategy achieving exactly $(\sqrt{W}, \sqrt{W})$. The reduction follows.

Lemma 24. *Given $v_1, \ldots, v_n \in \mathbb{Z}^+$, and $W = \prod_{i=1}^n \frac{1}{v_i}$, let $\epsilon < \frac{1}{4}\sqrt{W}$. If there exists a strategy τ such that for $i = 1, 2$, $p_i^\tau = \sqrt{W} + \delta_i$ for some $\delta_1, \delta_2 \in [-\epsilon, \epsilon]$, then there is a pure strategy achieving probabilities $(\sqrt{W}, \sqrt{W})$.*

Proof. Consider any σ with value vector $(\sqrt{W} + \delta_1, \sqrt{W} + \delta_2)$. If we write σ as the linear combination of pure strategies as $\sigma = \sum_{i=1}^n \lambda_i \pi_i$, we get

$$\lambda_1 w_1 + \ldots + \lambda_n w_n = \sqrt{W} + \delta_1,$$
$$\lambda_1 W/w_1 + \ldots + \lambda_n W/w_n = \sqrt{W} + \delta_2,$$

where $w_i = p_1^{\pi_i}$, and $W/w_i = p_2^{\pi_i}$. By dividing the second equation by W, and distributing the right hand side in the linear combination in both lines, we rewrite this as

$$\lambda_1(w_1 - \sqrt{W}) + \ldots + \lambda_n(w_n - \sqrt{W}) = \delta_1,$$
$$\lambda_1(\tfrac{w_1-\sqrt{W}}{w_1\sqrt{W}}) + \ldots + \lambda_n(\tfrac{w_n-\sqrt{W}}{w_n\sqrt{W}}) - \tfrac{\delta_2}{W} = 0.$$

Towards a contradiction, assume that $w_i \neq \sqrt{W}$ for all i. Define $P \subsetneq \{1, \ldots, n\}$, the set of i such that $w_i - \sqrt{W} > 0$, and let $N = \{1, \ldots, n\} \setminus P$. For all $i \in P$, we have $1/w_i \geq \sqrt{1/W} + 1$ which means $w_i \geq \frac{\sqrt{W}}{1-\sqrt{W}}$. For $i \in N$, we similarly obtain

$w_i \leq \frac{\sqrt{W}}{1+\sqrt{W}}$. We obtain that for any $i \in P$, $\frac{1}{w_i\sqrt{W}} \leq \frac{1-\sqrt{W}}{W}$, and for any $i \in N$, $\frac{1}{w_i\sqrt{W}} \geq \frac{1+\sqrt{W}}{W}$. We rewrite

$$\begin{array}{l}\sum_{i\in P} \lambda_i(w_i - \sqrt{W}) - \sum_{i\in N} \lambda_i(\sqrt{W} - w_i) = \delta_1 \in [-\epsilon, \epsilon], \\ \sum_{i\in P} \lambda_i \frac{w_i-\sqrt{W}}{w_i\sqrt{W}} + \frac{\delta_2}{W} \sum_{i\in P} \lambda_i = \sum_{i\in N} \lambda_i \frac{\sqrt{W}-w_i}{w_i\sqrt{W}} - \frac{\delta_2}{W} \sum_{i\in N} \lambda_i.\end{array} \tag{2}$$

We have

$$\begin{array}{l}\frac{1+\sqrt{W}}{W} \sum_{i\in N} \lambda_i(\sqrt{W} - w_i) - \frac{\delta_2}{W} \sum_{i\in N} \lambda_i \\ \leq \sum_{i\in N} \lambda_i \frac{\sqrt{W}-w_i}{w_i\sqrt{W}} - \frac{\delta_2}{W} \sum_{i\in N} \lambda_i \\ = \sum_{i\in P} \lambda_i \frac{w_i-\sqrt{W}}{w_i\sqrt{W}} + \frac{\delta_2}{W} \sum_{i\in P} \lambda_i \\ \leq \frac{1-\sqrt{W}}{W} \sum_{i\in P} \lambda_i(w_i - \sqrt{W}) + \frac{\delta_2}{W} \sum_{i\in P} \lambda_i.\end{array}$$

It folows that $(1 + \sqrt{W})\alpha_N - (1 - \sqrt{W})\alpha_P \leq \delta_2$, where $\alpha_P = \sum_{i\in P} \lambda_i(w_i - \sqrt{W})$ and similarly for α_N. We must have $(\alpha_N - \alpha_P) + \sqrt{W}(\alpha_N + \alpha_P) \leq \epsilon$. Because $\alpha_P - \alpha_N \in [-\epsilon, \epsilon]$, we have $\sqrt{W}(\alpha_N + \alpha_P) \leq 2\epsilon$. But we have that $|w_i - \sqrt{W}| \geq \frac{\sqrt{W}}{1+\sqrt{W}}$ so $\alpha_N + \alpha_P \geq \frac{\sqrt{W}}{1+\sqrt{W}}$. It follows that $\frac{W}{1+\sqrt{W}} \leq 2\epsilon$, which is a contradiction. ☐

7.2 Fixed-Memory Strategies

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).

This case will be used, in the next section, to derive an approximation algorithm for the general problem.

We start by analyzing the case of MDPs. Given an MDP $M = (S, A, \delta, r)$, and target set T, consider a subset S^{no} of states and $S^? = S \setminus (S^{\text{no}} \cup T)$. We will write an equation to solve the reachability problem as follows. For a starting state s_0, and desired reachability probability λ, we define the following equation with unknowns $x_s, p_{s,a}$ for all $s \in S^?$, $a \in A(s)$.

$$\begin{array}{l}\forall s \in S^{\text{no}}, x_s = 0, \\ \forall s \in T, x_s = 1, \\ \forall s \in S^?, x_s = \sum_{a\in A(s)} p_{s,a} \sum_{t\in S} \delta(s, a, t)x_t, \\ \forall s \in S, \sum_{a\in A(s)} p_{s,a} = 1, \\ \forall s \in S, a \in A(s), p_{s,a} \geq 0, \\ x_{s_0} \geq \lambda\end{array} \tag{3}$$

For any solution $(\bar{x}, \bar{p})$ of (3), let us denote by $\sigma_{\bar{p}}$ the strategy defined by $\sigma_{\bar{p}}(s, a) = p_{s,a}$. Let us also denote by $M^{\bar{p}}$ the Markov chain obtained from M by fixing the probability of each action a from s to $p_{s,a}$.

Lemma 25. *Consider any $S^{no} \subseteq S$ and any solution $\bar{x}, \bar{p}$ of* (3)*. If all states s of $M^{\bar{p}}$ with zero probability of reaching T belongs to S^{no}, then $x_s = \mathbb{P}^{\sigma_{\bar{p}}}_{M,s}[\mathsf{Reach}(T)]$.*

Conversely, for any stationary strategy σ, such that $\mathbb{P}^{\sigma}_{M,s_0}[\mathsf{Reach}(T)] \geq \lambda$, there exists a subset $S^{no} \subseteq S$ such that $x_s = \mathbb{P}^{\sigma}_{M,s}[\mathsf{Reach}(T)]$ and $p_{s,a} = \sigma(s,a)$ are the unique solution of (3)*.*

Proof. Fix any solution $(\bar{x}, \bar{p})$ of (3), and assume that all states s with a probability of 0 of reaching T satisfy $x_s \in S^{\text{no}}$. Then $\bar{x}$ is the solution of the equation obtained by fixing $\bar{p}$. But this equation has a unique solution which gives the reachability probabilities from each state (see *e.g.* [1, Theorem 10.19]).

Conversely, given a stationary strategy σ, we can define S^{no} as the set of states from which no path leads to T in the Markov chain M^{σ}, and by fixing the probabilities $p_{s,a} = \sigma(a \mid s)$ in (3), the unique solution is the vector of reachability probabilities. □

We now adapt (3) to MEMDPs and prove the following theorem.

Theorem 26. *The quantitative reachability and safety problems for K-memory strategies can be solved in polynomial space in K and in the size of M.*

Proof. We give the proof for reachability objectives. The case of safety is very similar and will be sketched.

For any MEMDP M, and given target states T, let us fix $S_i^{\text{no}} \subseteq S$, for each M_i. Given K, define the set $M = \{1, \ldots, K\}$ of *memory elements*, and fix an initial memory element $m_0 \in M$. Given desired reachability probabilities α_1, α_2 from state s_0, we write the following equation $E(S_1^{\text{no}}, S_2^{\text{no}})$.

$$
\begin{aligned}
&\forall s \in S_1^{\text{no}}, m \in M, x_{s,m} = 0,\\
&\forall s \in T, m \in M, x_{s,m} = 1,\\
&\forall s \in S_1^{?}, m \in M, x_{s,m} = \textstyle\sum_{a \in A(s), m' \in M} \sum_{t \in S} p_{s,m}(a,m')\delta_1(s,a,t)x_{t,m'},\\
&\forall s \in S_2^{\text{no}}, m \in M, y_{s,m} = 0,\\
&\forall s \in T, m \in M, y_{s,m} = 1,\\
&y_{s,m} = \textstyle\sum_{a \in A(s), m' \in M} \sum_{t \in S} p_{s,m}(a,m')\delta_2(s,a,t)y_{t,m'},\\
&\forall s \in S, m \in M, \textstyle\sum_{a \in A(s), m' \in M} p_{s,m}(a,m') = 1,\\
&\forall s \in S, a \in A(s), m, m' \in M, p_{s,m}(a,m') \geq 0,\\
&x_{s_0,m_0} \geq \alpha_1, y_{s_0,m_0} \geq \alpha_2.
\end{aligned}
\tag{4}
$$

The equation consists in embedding the memory in the MDPs. Each unknown $p_{s,m}(a,m')$ corresponds to the probability of choosing action a and changing memory to m' given state s and memory m. Thus $\sum_{m' \in M} p_{s,m}(a,m')$ is the probability of choosing action a at s, m.

Now polynomial space procedure proceeds as follows. We first guess the sets $S_1^{\text{no}}, S_2^{\text{no}}$, write the equation $E(S_1^{\text{no}}, S_2^{\text{no}})$, and solve it in deterministic polynomial space. We then check, for each $i = 1, 2$, whether all states s from which the probability of reaching T is 0 belongs to S_i^{no}. We accept if this is the case, and reject otherwise.

The correctness follows from Lemma 25. In fact, if there is a stationary strategy achieving probabilities α_1 and α_2 and s_0, then there exist the sets $S_1^{\text{no}}, S_2^{\text{no}}$ of 0-probability states, and for this guess (4) has a solution obtained by fixing $p_{s,a} = \sigma(a \mid$

s), and where x_s is the probability achived in M_1 at s, and y_s at M_2. Therefore the procedure accepts. If there is no such strategy, then for all guesses, either the desired probabilities do not satisfy the lower bounds, or one of the sets S_i^{no} does not contain all 0-probability states.

The problem can be solved similarly for safety properties. In fact, the events of avoiding T and reaching T are complementary. Equation (3) and Lemma 25 can be adapted for safety objectives by simply requiring $x_{s_0} \leq \lambda$, which means that the safety property holds with probability at least $1 - \lambda$ in Equation (4). $\square$

7.3 Approximation Algorithm

We now show that considering finite-memory strategies are hardly restrictive, in the sense that they can be used to approximately achieve the value. We also give a memory bound that is sufficient to approximate the value by any given ϵ.

Theorem 27. *For any MEMDP M with trivial DECs, 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 $\eta = \min\{|\delta_1(s, a, s') - \delta_2(s, a, s')| \mid s, a, s' \text{ s.t. } \delta_1(s, a, s') \neq \delta_2(s, a, s')\}$.*

of Lemma 27. By Definition 11 and Corollary 15, we assume that M has only trivial DECs. Consider an arbitrary strategy σ for M. Define $\eta = \min\{|\delta_1(s, a, s') - \delta_2(s, a, s')| \mid s, a, s' \text{ s.t. } \delta_1(s, a, s') \neq \delta_2(s, a, s')\}$. We call the pair (s, a) *distinguishing* if for some s', $\delta_1(s, a, s') \neq \delta_2(s, a, s')$. Let us fix $K = 2\frac{\log(1/\epsilon)}{\eta^2}$. Let p denote the smallest nonzero probability in M and $q = p^{|S|}$.

Strategy σ' is defined identically to σ on all histories up to length $L = l|S|$, where $l \geq \left(\frac{2|S||A|}{p^{|S|}\eta^2}\right)^2 \log^3(1/\epsilon)$. Note that L is exponential, so the memory requirement is doubly exponential. Upon arrival to a DEC (thus, trivial and absorbing) it switches to a memoryless strategy. On any other history $h_1 \ldots h_L$, we distinguish cases:

Assume there is a distinguishing pair that was seen at least K times in h, and consider (s, a) the first such pair. Let us write $d_i = \delta_i(s, a, s')$ for some s' with $d_1 \neq d_2$. Assume $d_i < d_{3-i}$ for some $i = 1, 2$. Define $c^L_{s,a}$ as the random variable denoting the number of occurences of (s, a) in a prefix of length L, and $c^L_{s,a,s'}$ the number of times the state s' was reached after (s, a). In σ', if $|\frac{c^L_{s,a,s'}}{c^L_{s,a}} d_i| < \frac{|d_1 - d_2|}{2}$, then we switch to the memoryless optimal strategy for M_i. If no distinguishing pair satisfies this condition for any $i = 1, 2$, then we switch to some arbitrary memoryless strategy. Strategy σ' is clearly finite-memory using $(|S| \cdot |A|)^L$ memory elements, for any choice of l.

First, let us show that conditioned on the event that some distinguishing pair was observed K times, the strategy σ' is ϵ-optimal. In fact, By Hoeffding's inequality, for an edge (s, a, s'), we have for any strategy τ,

$$\mathbb{P}^{\tau}_{M_i,s}\left[\left|\frac{c^L_{s,a,s'}}{c^L_{s,a}} - d_i\right| \geq \left|\frac{d_2 - d_1}{2}\right| \mid c^L_{s,a} \geq K\right] \leq e^{-2K\frac{d_2-d_1}{2}^2} \leq \epsilon,$$

which means that σ' will switch to the optimal strategy for M_i from with probability at least $1-\epsilon$.

Let us denote by T_K^L the event that $\exists(s,a), c_{s,a}^L = K$, and D^L the event that some DEC (therefore, trivial and absorbing) is reached. The rest of the proof consists in showing that with high probability either T_K^L occurs or the play is stuck in some absorbing state, and in any such history σ' performs as good as σ up to ϵ.

Either T_K^L or a DEC. We will show that either T_K^L or D^L occurs with probability $1-\epsilon$.

Let X_j denote the random variable giving the state at j-th step, and A_j the j-th action. For any history h, let $y(h)$ denote the number of states of h belonging to a DEC + the number of distinguishing actions in h. We show that, under any strategy τ, and for any state s, $i=1,2$, $\mathbb{P}^\tau_{M_i,s}[y(X_1A_1\dots A_{|S|-1}X_{|S|}) \geq 1] \geq q$. To prove this, we first write τ as a linear combination of strategies that are deterministic in the first $|S|$ steps: $\tau = \sum_i \lambda_i \pi_i$ for $(\pi_i)_i$ a finite family of strategies that are pure in the first $|S|$ steps, and $(\lambda_i)_i$ such that $\sum_i \lambda_i = 1$. We have that $\mathbb{P}^\tau_{M_i,s}[y(X_1A_1\dots A_{|S|-1}X_{|S|})] = \sum_j \lambda_j \mathbb{P}^{\pi_j}_{M_i,s}[y(X_1A_1\dots A_{|S|-1}X_{|S|}) \geq 1]$. We will prove that for each π_j,

$$\mathbb{P}^{\pi_j}_{M_i,s}[y(X_1A_1\dots A_{|S|-1}X_{|S|}) \geq 1] \geq q$$

. We consider the unfolding of depth $|S|$ from state s under strategy π_j. If this unfolding contains a state t of a DEC, then the path from s to t has probability at least q under strategy π_j since it is deterministic in the first $|S|$ steps, and the result follows. If the unfolding contains a distinguishing action, then it will be taken similarly with probability at least q. Otherwise, assume the unfolding contains no DEC or distinguishing action. But in this case, if we cut each branch whenever a state is visited twice, we obtain an end-component in M_i. Since no action is distinguishing, this is a non-distinguishing double end-component, which is a contradiction.

It follows that $\mathbb{E}^\tau_{M_i,s}[y(X_1A_1\dots A_{|S|-1}X_{|S|})] \geq q$ for any state s and any strategy τ. We factorize a given history of length L in to factors of length $|S|$. Let Y_j be the random variable denoting $y(h_{(j-1)|S|+1\dots j|S|})$. We just showed that $\mathbb{E}^\tau_{M_i,s}[Y_j] \geq q$ for any strategy τ, state s and $j = 1\dots l$. Let $Y = \sum_{j=1}^l Y_j$. We use Hoeffding's inequality to write

$$\mathbb{P}^\tau_{M_i,s}[Y \leq \mathbb{E}[Y] - t] \leq e^{-\frac{t^2}{2l|S|^2}},$$

for any $t > 0$, since $\left|\mathbb{E}^\tau_{M_i,s}[Y_j]\right| \leq 2|S|$. We get that $\mathbb{P}^\tau_{M_i,s}[Y \leq lq - t] \leq e^{-\frac{t^2}{2l|S|^2}}$ since $lq \leq \mathbb{E}[Y]$. We would like to obtain that $\mathbb{P}^\tau_{M_i,s}[Y \leq |S|\cdot|A|\cdot K] \leq \epsilon$, which means that with probability at least $1-\epsilon$, either T_K^L or D^L holds. Therefore, in the above equation, we require $e^{-\frac{t^2}{2l|S|^2}} \leq \epsilon$, which means

$$\frac{t^2}{l} \geq 2\log(1/\epsilon)|S|^2, \tag{5}$$

and we let $t = lq - |S||A|K$. To get (5), it suffices to ensure $\frac{(lq-|S||A|K)^2}{l} \geq 2\log(1/\epsilon)|S|^2$, which holds for our choice of l.

End of the proof We write $\mathbb{P}^{\sigma'}_{M_i,s}[\phi]$ as

$$\mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid T^L_K \vee D^L]\mathbb{P}^{\sigma'}_{M_i,s}[T^L_K \vee D^L] + \mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid \neg T^L_K \wedge \neg D^L]\mathbb{P}^{\sigma'}_{M_i,s}[\neg T^L_K \wedge \neg D^L]$$

We clearly have $\mathbb{P}^{\sigma}_{M_i,s}[T^L_K \vee D^L] = \mathbb{P}^{\sigma'}_{M_i,s}[T^L_K \vee D^L]$, and we showed above that $\mathbb{P}^{\sigma'}_{M_i,s}[T^L_K \vee D^L] \geq 1-\epsilon$. Thus, using the same decomposition, $\mathbb{P}^{\sigma}_{M_i,s}[\phi \mid T^L_K \vee D^L] \geq \mathbb{P}^{\sigma}_{M_i,s}[\phi] - \epsilon$.

We will show that $\mathbb{P}^{\sigma'}_{M_i,s}[\phi \wedge (T^L_K \vee D^L)] \geq (1-\epsilon)\mathbb{P}^{\sigma}_{M_i,s}[\phi \wedge (T^L_K \vee D^L)]$, which implies $\mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid T^L_K \vee D^L] \geq (1-\epsilon)\mathbb{P}^{\sigma}_{M_i,s}[\phi \mid T^L_K \vee D^L]$. But let us first show how we conclude. Because $\mathbb{P}^{\sigma'}_{M_i,s}[\phi] \geq \mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid T^L_K \vee D^L]\mathbb{P}^{\sigma'}_{M_i,s}[T^L_K \vee D^L]$, combining with the above inequality, it follows

$$\begin{aligned} \mathbb{P}^{\sigma'}_{M_i,s}[\phi] & \geq (1-\epsilon)^2(\mathbb{P}^{\sigma}_{M_i,s}[\phi] - \epsilon) \\ & \geq \mathbb{P}^{\sigma}_{M_i,s}[\phi] - 3\epsilon, \end{aligned}$$

as desired.

We write $\mathbb{P}^{\sigma'}_{M_i,s}[\phi \wedge (T^L_K \vee D^L)] = \mathbb{P}^{\sigma'}_{M_i,s}[\phi \wedge T^L_K] + \mathbb{P}^{\sigma'}_{M_i,s}[\phi \wedge D^L \wedge \neg T^L_K]$. We have $\mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid D^L \wedge \neg T^L_K] = \mathbb{P}^{\sigma}_{M_i,s}[\phi \mid D^L \wedge \neg T^L_K]$ for both $i = 1, 2$ since the history ends in an absorbing state. It follows that $\mathbb{P}^{\sigma'}_{M_i,s}[\phi \wedge D^L \wedge \neg T^L_K] = \mathbb{P}^{\sigma}_{M_i,s}[\phi \wedge D^L \wedge \neg T^L_K]$. We now show that $\mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid T^L_K] \geq (1-\epsilon)\mathbb{P}^{\sigma}_{M_i,s}[\phi \mid T^L_K]$ which implies similarly $\mathbb{P}^{\sigma'}_{M_i,s}[\phi \wedge T^L_K] \geq (1-\epsilon)\mathbb{P}^{\sigma}_{M_i,s}[\phi \wedge T^L_K]$ since $\mathbb{P}^{\sigma'}_{M_i,s}[T^L_K] = \mathbb{P}^{\sigma}_{M_i,s}[T^L_K]$. Let $S^L_K(i)$ denote the event that for the first distinguishing pair (s, a) that appears K times in the prefix of length L, $|\frac{c^L_{s,a,s'}}{c^L_{s,a}} - d_i| \leq |\frac{d_1-d_2}{2}|$. We have

$$\begin{aligned} \mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid T^L_K] & = \mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid S^L_K(i) \wedge T^L_K]\mathbb{P}^{\sigma'}_{M_i,s}[S^L_K(i) \mid T^L_K] + \\ & \quad \mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid \neg S^L_K(i) \wedge T^L_K]\mathbb{P}^{\sigma'}_{M_i,s}[\neg S^L_K(i) \mid T^L_K]. \end{aligned}$$

For any $i = 1, 2$, we have $\mathbb{P}^{\sigma'}_{M_i,s}[\neg S^L_K(i) \mid T^L_K] \leq \epsilon$ as we showed above, so $\mathbb{P}^{\sigma'}_{M_i,s}[S^L_K(i) \mid T^L_K] \geq 1-\epsilon$, and $\mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid S^L_K(i) \wedge T^L_K] \geq \mathbb{P}^{\sigma}_{M_i,s}[\phi \mid S^L_L(i) \wedge T^L_K]$ since σ' switches to the optimal strategy for M_i. This shows that $\mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid T^L_K] \geq (1-\epsilon)\mathbb{P}^{\sigma'}_{M_i,s}[\phi \mid T^L_K]$. ☐

Combining Lemmas 26 and 27, we derive an approximation algorithm:

Theorem 28. *There is a procedure that works in $O(N \cdot |M|)$ space solving the ϵ-gap problem for quantitative reachability in MEMDPs. Moreover, whenever the procedure answers YES, there exists a strategy σ such that $\forall i = 1, 2, \mathbb{P}^{\sigma}_{M_i,s}[\mathsf{Reach}(T)] \geq \alpha_i - \epsilon$.*

of Lemma 28. We compute N given by Lemma 27, which is doubly exponential in input, and apply Lemma 26 for N-memory strategies and target probabilities $\alpha_1 - \epsilon$ and $\alpha_2 - \epsilon$. We solve the equation in polynomial space in the equation size, and answer yes if, and only if there is a solution.

By Lemma 27, if there exists a strategy achieving (α_1, α_2), there is a N-memory strategy achieving $(\alpha_1 - \epsilon, \alpha_2 - \epsilon)$. So the procedure will answer yes. If no strategy achieves $(\alpha_1 - \epsilon, \alpha_2 - \epsilon)$, then, in particular, no finite-memory strategy achieves this vector, and the procedure will answer no.

Last, observe that whenever the procedure answers yes, there exists a finite-memory strategy achieving $(\alpha_1 - \epsilon, \alpha_2 - \epsilon)$. $\square$

In our case, 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. Notice that we do not have false positives; when the procedure answers positively, the probabilities are achieved up to ϵ.

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 29. *The ϵ-gap problem for MEMDPs is NP-hard.*

Proof. We prove that the Product-Partition problem is polynomial-time reducible to the ϵ-gap problem for quantitative reachability in MEMDPs. Consider an instance given by $v_1, \ldots, v_n, A$ and consider the MEMDP M constructed above with starting state s_1 and target $\{s_n\}$. Let W and ϵ be given by the above lemma. Note that W and ϵ can be computed in polynomial time.

If Product-Partition has a solution, then there exists a pure strategy in M achieving $(\sqrt{W}, \sqrt{W})$ and the ϵ-gap instance is positive. If there is Product-Partition has no solution, then there is no pure strategy achieving this vector. Therefore, by Lemma 24, there is no strategy achieving at least $(\sqrt{W} - \epsilon, \sqrt{W} - \epsilon)$ componentwise. Thus, the ϵ-gap instance is negative. $\square$

8 Safety and Parity Objectives

8.1 The Almost-sure Case

We consider safety and parity objectives, building on techniques developed for reachability. Recall that almost-sure and sure safety coincide in MEMDPs. The equivalence of these with limit-sure safety is less trivial, and follows from Lemma 17:

Lemma 30. *Limit-sure safety is equivalent to almost-sure safety in MEMDPs.*

Under Assumption 7, safety is a special case of parity objectives; we rely on algorithms for parity to decide almost-sure safety objectives. For quantitative safety, the results of the previous section can be adapted without difficulty, but we omit the details.

Our first result is a polynomial-time algorithm for almost sure parity objectives.

By Lemma 1, we know that for MDPs with parity objectives, inside end-components, the value of a Parity objective is either 0 or 1, and in the latter case a memoryless strategy which only depends on the support of the distributions exists. Thus, let us call an end-component D Φ-*winning* if there exists a strategy inside D that satisfies Φ.

We denote by R^{Φ} the set of revealed states from which Φ surely holds.

Lemma 31. *For any MEMDP M, state s, and parity objective Φ, Algorithm 2 decides whether there exists a strategy achieving Φ almost surely in M, and computes a witnessing memoryless strategy.*

Input: MEMDP M, $s \in S$, parity objective Φ

$U := \big(\mathsf{AS}(M_1, \Phi) \cap \mathsf{AS}(M_2, \Phi)\big) \cup R^{\Phi}$;

M' := Sub-MEMDP of M induced by states s' s.t. $\mathsf{Val}^*_{\mathsf{Safe}(U)}(\cup M, s') = 1$;

T_i:= Set of states of Φ-winning MECs of M'_i;

if $\exists \sigma, \forall i = 1, 2, \mathbb{P}^{\sigma}_{M'_i,s}[\mathsf{Reach}(T_1 \cup T_2)] = 1$ **then**

  σ' := Modify σ as follows. At any state $s \in T_1$ (resp. $s \in T_2 \setminus T_1$), if D denotes the MEC of M_1 (resp. M_2) which contains s, switch to a memoryless strategy winning for Φ compatible with D;

  Return σ';

else

  Return NO;

end

Algorithm 2: Almost-sure parity algorithm for MEMDPs

Proof. Consider an instance M, s and Φ for which the algorithm answers positively. Under the returned strategy σ', some state s' from $T_1 \cup T_2$ is visited almost surely for the first time in M_i. If σ' switches to an optimal strategy for M'_i (that is, either $i = 1$ and $s' \in T_1$, or $i = 2$ and $s' \in T_2 \setminus T_1$), then Φ holds almost surely in M_i by definition. Otherwise, $s' \in D$ for a Φ-winning MEC D of M'_{3-i} and σ' switches to an optimal strategy for M'_{3-i} that stays in D. Let $D'_1, \ldots, D'_m$ be the set of all end-components included in D such that $\mathbb{P}^{\sigma'}_{M'_{3-i},s'}[\mathsf{Inf} = D'_j] > 0$. First, observe that by Assumption 6, $R^{\Phi} \cup \bigcup_{j=1}^{m} D'_j$ is reached almost surely in M'_i under σ' from s', since σ' is compatible with D in M'_{3-i}. Moreover, $\mathbb{P}^{\sigma'}_{M'_i,s'}[\mathsf{Inf} \in \{D'_1, \ldots, D'_m\} \cup R^{\Phi}] = 1$. If some D'_j is a DEC, then σ' also almost surely satisfies Φ in M'_i from D'_j by Lemma 1. Otherwise, some action a from some state t of D'_j has a different support in M'_i and M'_{3-i}. Because D'_j is an end-component for M'_{3-i}, we have $\mathsf{Supp}(\delta_{3-i}(t, a)) \subsetneq \mathsf{Supp}(\delta_i(t, a))$ since otherwise D'_j would contain an absorbing state different than t (by Assumption 6), which is in contradiction with the fact that it is an end-component in M'_{3-i}. Therefore, starting at D'_j in M'_i, under σ', the play almost surely leaves D'_j for a i-revealed state. By definition of M' such a state is in R^{Φ}. Thus, $\mathbb{P}^{\sigma'}_{M_i,s'}[\Phi] = 1$.

Conversely, assume that there exists τ such that $\mathbb{P}^{\tau}_{M_i,s}[\Phi] = 1$ for all $i = 1, 2$. Observe that $\mathbb{P}^{\tau}_{M_i,s}[\mathsf{Safe}(U)] = 1$ since otherwise for some $i = 1, 2$, we reach a state that is not almost surely winning. Strategy τ is therefore compatible with M'_i and $s \in M'_i$. Recall that a strategy satisfies a parity condition almost surely in an MDP if, and only if the set of winning MECs is reached almost surely. Hence, we must have $\forall i = 1, 2, \mathbb{P}^{\tau}_{M'_i,s}[\mathsf{Reach}(T_i)] = 1$, so in particular $\forall i = 1, 2, \mathbb{P}^{\tau}_{M'_i,s}[\mathsf{Reach}(T_1 \cup T_2)] = 1$, and the algorithm answers positively. □

This yields the following theorem.

Theorem 32. *The almost-sure parity problem is decidable in polynomial time.*

8.2 The Quantitative Case: Reduction to Reachability

Our second result is a polynomial-time reduction from the quantitative parity problem to the quantitative reachability problem which preserves value vectors. It follows 1) a polynomial-time algorithm for the limit-sure parity problem, 2) and that any algorithm for solving the quantitative reachability problem can be used to solve the quantitative parity problem. In particular, results of Section 7 applies to parity objectives.

The idea of the reduction is similar to previous constructions. We modify $\hat{M}$ by adding new transitions from each MEC D of each M_i to fresh absorbing states with probability equal to the probability of winning from D in M_i.

Definition 33. *Given a MEMDP M, we define $\bar{M} = (\bar{S}, \bar{A}, \bar{\delta}_1, \bar{\delta}_2)$ by modifying $\hat{M}$ as follows. For any $i = 1, 2$, non-trivial MEC D of $\hat{M}_i$, and state $s \in D$, we add an action a_D from s. In $\bar{M}_i$, a_D leads to a fresh absorbing state t^0_D with even parity with probability $\mathsf{Val}^*_{\mathcal{P}}(M_i, s)$, and to a fresh absorbing state t^1_D with odd parity with remaining probability. In $\bar{M}_{3-i}$, it leads to a losing absorbing state t^1_D. Let $\bar{\Phi}$ be the reachability objective with targets the absorbing states of $\cup\bar{M}$ with even parity.*

Observe that the set of absorbing states of $\cup\bar{M}$ with even parity is exactly $\{W_D \mid D \text{ DEC of } \hat{M}\} \cup \{t^0_D \mid \exists j = 1, 2, D \in \mathcal{G}(\bar{M}_j)\}$. For any state s of M, let $\bar{\mathcal{A}}(s)$ denote the state in $\bar{M}$ to which it is mapped by our construction: for any s belonging to a MDEC D, $\bar{\mathcal{A}}(s) = s_D$, and $\bar{\mathcal{A}}(s) = s$ otherwise. Note that $\bar{M}$ can be constructed in polynomial time since MECs can be computed in polynomial time.

We will prove that achieving a pair of satisfaction probabilities for a parity objective $\mathcal{P}_p$ in M is equivalent to achieving the same probabilities for the reachability objective $\bar{\Phi}$ in $\bar{M}$.

We start with two simple technical lemmas. Let us denote by $\mathcal{G}(\bar{M}_j)$ the set of MECs of $\bar{M}_j$ that are not DECs. The following lemma gives a classification of the MECs of M_i with respect to $\bar{M}_i$.

Lemma 34. *Let D be an end-component of M_i which is not a DEC. Then, either D contains a distinguishing DEC, or $D \subseteq \bar{\mathcal{A}}^{-1}(E)$ for some $E \in \mathcal{G}(\bar{M}_i)$.*

Proof. Assume that D is not a DEC and does not contain distinguishing DECs. Notice that D might contain non-distinguishing DECs. By construction $\bar{\mathcal{A}}(D)$ is an end-component in $\bar{M}_i$: it is δ_i-closed and strongly connected. We have $D \subseteq \bar{\mathcal{A}}^{-1}(\bar{\mathcal{A}}(D))$. ☐

The following lemma is an adaptation of Lemma 13 to paths in M_{3-j} that stay in the preimage of a MEC of $\bar{M}_j$.

Lemma 35. *For any MEMDP M, and $\epsilon > 0$, there exists K such that for any $j = 1, 2$, $D \in \mathcal{G}(\hat{M}_j)$, and any history $h \in \mathcal{H}(\bar{M})$ which contains a factor of length K compatible with D, $\mathbb{P}^{\tau}_{M_{3-j},s}[\mathsf{red}^{-1}(h)] \leq \epsilon$ for any strategy τ.*

Proof. For any $\epsilon > 0$, fix K as in Lemma 13. Fix any $D \in \mathcal{G}(\hat{M}_j)$, and history $h \in \mathcal{H}(\hat{M})$ of length K compatible with D. We know that $\mathbb{P}^{\tau}_{\hat{M}_{3-j},h_1}[h] \leq \epsilon$ by Lemma 13. History h does not contain states $\mathcal{T}$ since otherwise it enters an absorbing state. It follows, from Lemma 14 that $\mathbb{P}^{\tau}_{M_{3-j},s}[\mathsf{red}^{-1}(h)] \leq \epsilon$. ☐

We can now prove the following direction.

Lemma 36. *Consider any MEMDP M, and parity condition $\mathcal{P}$. For any state s, strategy σ, and $\epsilon > 0$, there exists a strategy $\bar\sigma$ such that $\mathbb{P}^{\bar\sigma}_{\bar M_i,\bar s}[\bar\Phi] \geq \mathbb{P}^{\sigma}_{M_i,s}[\mathcal{P}] - \epsilon$.*

Proof. Let $\bar\sigma$ as defined in Lemma 14. We define $\bar\sigma'$ from $\bar\sigma$ as follows. For any ϵ let K be as defined in Lemma 35. For any $j = 1, 2$, and $D \in \mathcal{G}(\bar M_j)$, define $\mathcal{D}^j_K(D)$ as the set of histories in $\mathcal{H}_{\mathcal{T}}(\bar M)$ whose suffix of length K is compatible with D, and such that no proper suffix contains a factor of length K compatible with any $D' \in \mathcal{G}(\bar M_1) \cup \mathcal{G}(\bar M_2)$. Let $\mathcal{D}^j_K$ denote the union of all $\mathcal{D}^j_K(D)$, and $\mathcal{D}_K = \mathcal{D}^1_K \cup \mathcal{D}^2_K$.

The following events are disjoint and occur almost surely in each $\bar M_j$ under $\bar\sigma'$:

$\bar E_1$: $\mathcal{D}_K \bar S^\omega$.

$\bar E_2$: $\mathcal{H}_{\mathcal{T}} \mathcal{T} \bar S^\omega \setminus \bar E_1$.

$\bar E_3$: $\mathcal{H}_{\mathcal{T}} \cdot a^{\$}_D \cdot \bar S^\omega \setminus \bar E_1$ for some non-distinguishing DEC D.

This follows from the fact that states and actions seen infinitely often in $\bar M_j$ under $\bar\sigma'$ is almost surely an end-component. If such an end-component is in $\mathcal{G}(\bar M_j)$ then $\bar E_1$ holds. If $\bar E_1$ does not hold, then any such end-component is a DEC, thus a trivial DEC. Because the DEC $\{t_D\}$ is only reachable if $\mathcal{D}_K$ occurs by definition of $\bar\sigma'$, any such end-component corresponds to a distinguishing or non-distinguishing DEC.

Similarly, the following events are disjoint and occur almost surely in each M_j under σ:

E_1 : $\mathsf{red}^{-1}(\mathcal{D}_K) S^\omega$.

E_2 : $\mathsf{red}^{-1}(\mathcal{H}_{\mathcal{T}}) \bar{\mathcal{A}}^{-1}(\mathcal{T}) \setminus E_1$.

E_3 : $\mathsf{red}^{-1}(\mathcal{H}_{\mathcal{T}}) D^\omega \setminus E_1$ for some non-distinguishing DEC D.

In fact, if the play stays in an end-component that belongs to $\bar{\mathcal{A}}^{-1}(E)$ for some $E \in \mathcal{G}(\bar M_j)$ then we are in E_1. If E_1 is false, then such an end-component either contains a distinguishing DEC, in which case E_2 holds almost surely, or it is a DEC, in which case either E_2 or E_3 holds almost surely.

By Lemma 35, we have $\mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(\mathcal{D}_K(D))] \leq \epsilon$ for $D \in \mathcal{G}(\bar M_{3-j})$. It follows that, for any $j = 1, 2$,

$$\begin{aligned}
\mathbb{P}^{\sigma}_{M_j,s}[\mathcal{P}] \quad & \geq \textstyle\sum_{H \in \mathsf{red}^{-1}(\mathcal{H}_{\mathcal{T}} \cdot \mathcal{T}) \setminus E_1} \mathbb{P}^{\sigma}_{M_j,s}[\mathcal{P} \mid \mathsf{red}^{-1}(H))] \mathbb{P}^{\sigma}_{M_j,s}[\mathsf{red}^{-1}(H)] \\
& \quad + \textstyle\sum_{H \in \mathsf{red}^{-1}(\mathcal{H}_{\mathcal{T}}) \setminus E_1, D \text{ non-dist.}} \mathbb{P}^{\sigma}_{M_j,s}[\mathcal{P} \mid H \cdot D^\omega] \mathbb{P}^{\sigma}_{M_j,s}[H \cdot D^\omega] \\
& \quad + \textstyle\sum_{H \in \mathsf{red}^{-1}(\mathcal{D}^j_K) S^\omega} \mathbb{P}^{\sigma}_{M_j,s}[\mathcal{P} \mid H] \mathbb{P}^{\sigma}_{M_j,s}[H]. \\
& \geq \mathbb{P}^{\sigma}_{M_j,s}[\mathcal{P}] - \epsilon.
\end{aligned}$$

We similarly write

$$\begin{aligned}
\mathbb{P}^{\bar\sigma'}_{\bar M_j,\bar s}[\bar\phi_j] \quad & \geq \textstyle\sum_{H \in \mathcal{H}_{\mathcal{T}} \cdot \mathcal{T} \setminus \bar E_1} \mathbb{P}^{\bar\sigma'}_{\bar M_j,\bar s}[\bar\phi_j \mid H] \mathbb{P}^{\bar\sigma'}_{\bar M_j,\bar s}[H] \\
& \quad + \textstyle\sum_{H \in \mathcal{H}_{\mathcal{T}} \setminus \bar E_1, D \text{ non-dist.}} \mathbb{P}^{\bar\sigma'}_{\bar M_j,\bar s}[\bar\phi_j \mid H \cdot a^{\$}_D (s^{\$}_D)^\omega] \mathbb{P}^{\bar\sigma'}_{\bar M_j,\bar s}[H \cdot a^{\$}_D (s^{\$}_D)^\omega] \\
& \quad + \textstyle\sum_{H \in \mathcal{D}^j_K} \mathbb{P}^{\bar\sigma'}_{\bar M_j,\bar s}[\bar\phi_j \mid H] \mathbb{P}^{\sigma}_{M_j,s}[H].
\end{aligned}$$

We will now compare the probability of winning in M and in $\bar{M}$ conditioned on events E_i and $\bar{E}_i$. First, note that by (1), and the definition of $\bar{\sigma}$, we have that for any $H \in \mathcal{H}_\mathcal{T} \cdot \mathcal{T} \cup \mathcal{H}_\mathcal{T} a_D^\$ (s_D^\$)^\omega \cup \mathcal{D}_K^j$, $\mathbb{P}^\sigma_{M_j,s}[\textsf{red}^{-1}(H)] = \mathbb{P}^{\bar{\sigma}'}_{\bar{M}_j,\bar{s}}[H]$. Let us show that conditioned on each of these events $\bar{M}_j$ achieves a higher or equal probability under $\bar{\sigma}$. For histories $\mathcal{H}_\mathcal{T} \cdot \mathcal{T}$ this is clear since $\bar{M}_j$ then reaches W_D with the optimal probability of winnig in M_j. For histories $\mathcal{H}_\mathcal{T} a_D^\$ (s_D)^\omega$, the probability achieved in $\bar{M}_j$ is exactly the probability of winning in M_j while staying inside D, so at least $\mathbb{P}^\sigma_{M_j,s}[\mathcal{P} \mid H \cdot D^\omega]$. Last, from histories $\mathcal{D}_K^j(D)$ with $D \in \mathcal{G}(\bar{M}_j)$, we reach in $\bar{M}_j$ with optimal probability of winning from a corresponding state in M_j, so at least $\mathbb{P}^\sigma_{M_j,s}[\mathcal{P} \mid h]$ for any $h \in \textsf{red}^{-1}(H)$. It follows that $\mathbb{P}^{\bar{\sigma}}_{\bar{M}_j,\bar{s}}[\bar{\phi}_j] \geq \mathbb{P}^\sigma_{M_j,s}[\mathcal{P}] - \epsilon$. ☐

To prove the converse, we need some additional lemmas.

Lemma 37. *Let D be a non-distinguishing DEC, and $\bar{\sigma}$ any strategy in $\bar{M}$. There exists a strategy τ such that for any history $hs_D(f_1a_1)s_D \ldots (f_ma_m)s'$, where (f_ia_i) is a pair of frontier state-action, and s' is a state outside of D,*

$$\mathbb{P}^\tau_{M_j,s_0}[\textsf{red}^{-1}(s_D(f_1a_1)s_D \ldots (f_ma_m)s') \mid h] = \mathbb{P}^{\bar{\sigma}}_{\bar{M}_j,\bar{s_0}}[s_D(f_1a_1)s_D \ldots (f_ma_m)s' \mid h],$$
$$\mathbb{P}^\tau_{M_j,s_0}[\textsf{red}^{-1}(s_D(f_1a_1)s_D \ldots (f_ma_m))D^\omega \mid h] = \mathbb{P}^{\bar{\sigma}}_{\bar{M}_j,\bar{s_0}}[s_D(f_1a_1)s_D \ldots (f_ma_m)a_D^\$ \mid h].$$

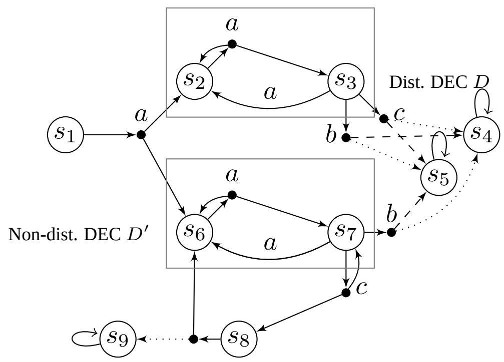
Proof. Consider any history $\textsf{red}^{-1}(s_D(f_1a_1)\ldots(f_{i-1}a_{i-1})s_D)$. let $p_{f,a}$ denote the probability that the pair of frontier state-action (fa) is taken for the first time under $\bar{\sigma}$ from history $s_D(f_1a_1)\ldots(f_{i-1}a_{i-1})s_D$. We define τ, by first choosing each pair (fa) with probability $p_{f,a}$, and then running a memoryless strategy that reaches state f almost surely, and once f is reached chooses a. With probability $1 - \sum_{f,a} p_{f,a}$, we run any strategy compatible with D. This is clearly the probability of $\bar{\sigma}$ of taking action $a_D^\$$. ☐

Lemma 38. *Consider any MEMDP M, and parity condition $\mathcal{P}$. For any state s, strategy $\bar{\sigma}$ for $\bar{M}$, and $\epsilon > 0$, one can compute σ such that $\mathbb{P}^\sigma_{M_i,s}[\mathcal{P}] \geq \mathbb{P}^{\bar{\sigma}}_{\bar{M}_i,\bar{s}}[\bar{\Phi}] - \epsilon$ for all $i = 1, 2$.*

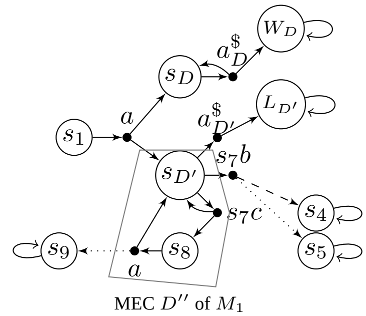
Proof. We define σ as follows. For any history $h_1 \ldots h_i \in \mathcal{H}_\mathcal{T}(\bar{M})$, such that $h_i \neq s_D$, and $a \in A(h_i)$, we let $\sigma(a \mid g) = \bar{\sigma}(a \mid h_1 \ldots h_i)$, for any $g \in \textsf{red}^{-1}(h_1 \ldots h_i)$. For any $g \in \textsf{red}^{-1}(h_1 \ldots h_i s_D)$ with $s_D \in \mathcal{T}$, we run the strategy given by Lemma 9 which achieves the objective with probability $1 - \epsilon$. For any $g \in \textsf{red}^{-1}(h_1 \ldots h_i s_D)$ where D is a non-distinguishing we switch to the strategy τ of Lemma 37 until D is left. Furthermore, at any history $g \in \textsf{red}^{-1}(h_1 \ldots h_i)$ with h_i belonging to some $D \in \mathcal{G}(\bar{M}_j)$, we switch to the optimal strategy for M_j from g with probability $\bar{\sigma}(a_D \mid h_1 \ldots h_i)$.

We can then easily prove the following correspondance between histories of $\bar{M}$ and that of M. For any history $h_1 \ldots h_i \in \mathcal{H}_\mathcal{T}(\hat{M})$, and action $a \in \hat{A}$,

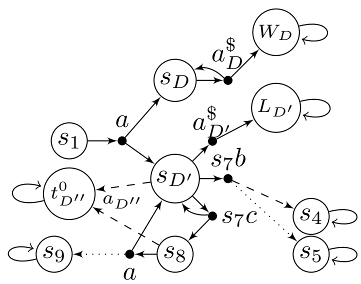
$$\begin{aligned} \mathbb{P}^{\bar{\sigma}}_{\bar{M}_j,\bar{s}}(h_1 \ldots h_i) &= \mathbb{P}^\sigma_{M_j,s}[\textsf{red}^{-1}(h_1 \ldots h_i)], \\ \mathbb{P}^{\bar{\sigma}}_{\bar{M}_j,\bar{s}}(h_1 \ldots h_i a) &= \mathbb{P}^\sigma_{M_j,s}[\textsf{red}^{-1}(h_1 \ldots h_i a)], \\ \mathbb{P}^{\bar{\sigma}}_{\bar{M}_j,\bar{s}}[h_1 \ldots h_i a_D^\$] &= \mathbb{P}^\sigma_{M_j,s}[\textsf{red}^{-1}(h_1 \ldots h_i D^\omega)]. \end{aligned} \tag{6}$$



(a) MEMDP M.



(b) MEMDP $\hat{M}$ where DECs were reduced.



(c) The MEMDP $\bar{M}$ obtained by adding the state $t^0_{D''}$ corresponding to the MEC D'' of M_1. The reachability objective is $\bar{\Phi} = \mathsf{Reach}(\{s_4, s_9, t^0_{D''}, W_D\})$.

Figure 5: We are given an MEMDP in Fig. 5a where we assume that D is a distinguishing DEC D, and D' is a non-distinguishing DEC (precise values of the probabilities do not matter). Distributions whose support differ in M_1 and M_2 are again shown in dashed or dotted lines. The parity function assigns 0 to s_4, s_9 and w_D, and 1 everywhere else.

Note that restricting histories to $\hat{M}$ and $\hat{A}$ only means that we exclude action a_D for MECs $D \in \mathcal{G}(\bar{M}_j)$.

The rest of the proof is done as in Lemma 36: We rewrite the probability of ensuring $\bar{\Phi}$ in $\bar{M}$.

$$\begin{aligned}
\mathbb{P}^{\bar{\sigma}'}_{\bar{M}_j,\bar{s}}[\bar{\phi}_j] \quad &\geq \textstyle\sum_{H \in \mathcal{H}_{\mathcal{T}} \cdot \mathcal{T} \setminus \bar{E}_1} \mathbb{P}^{\bar{\sigma}'}_{\bar{M}_j,\bar{s}}[\bar{\phi}_j \mid H] \mathbb{P}^{\bar{\sigma}'}_{\bar{M}_j,\bar{s}}[H] \\
&\quad + \textstyle\sum_{H \in \mathcal{H}_{\mathcal{T}} \setminus \bar{E}_1, D \text{ non-dist.}} \mathbb{P}^{\bar{\sigma}'}_{\bar{M}_j,\bar{s}}[\bar{\phi}_j \mid H \cdot a^{\$}_D (s^{\$}_D)^\omega] \mathbb{P}^{\bar{\sigma}'}_{\bar{M}_j,\bar{s}}[H \cdot a^{\$}_D (s^{\$}_D)^\omega] \\
&\quad + \textstyle\sum_{H \in \mathcal{D}^j_K} \mathbb{P}^{\bar{\sigma}'}_{\bar{M}_j,\bar{s}}[\bar{\phi}_j \mid H] \mathbb{P}^{\sigma}_{M_j,s}[H].
\end{aligned}$$

and show that conditioned on each above event, the probability of winning in M is at least as high as the expectation in $\bar{M}$, up to ϵ. On histories $H \in \mathcal{H}_{\mathcal{T}} \cdot \mathcal{T}$ this follows by Lemma 9. On histories that end with D^ω for non-distinguishing components D, the probability of winning in $\bar{M}_j$ is the optimal probability of winning in M_j from D, which is achieved by σ. Last, the probability of winning conditioned on $\mathcal{D}^j_K$ is equal to the optimal probability of winning in M_j from the current state by construction, and this is the probability achieved from such histories in M by definition of σ. $\square$

We summarize the result we proved in the following theorem.

Theorem 39. *The quantitative parity problem is polynomial-time reducible to the quantitative reachability problem. The limit-sure parity problem is in polynomial time.*

References

[1] C. Baier and J.-P. Katoen. *Principles of model checking*. MIT Press, 2008.

[2] J. Canny. Some algebraic and geometric computations in pspace. In *STOC'88*, STOC '88, pp. 460–467, New York, NY, USA, 1988. ACM.

[3] K. Chatterjee, M. Chmelik, and M. Tracol. What is decidable about partially observable markov decision processes with omega-regular objectives. In *CSL*, vol. 23 of *LIPIcs*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013.

[4] T. Chen, T. Han, and M. Z. Kwiatkowska. On the complexity of model checking interval-valued discrete time markov chains. *Inf. Process. Lett.*, 113(7):210–216, 2013.

[5] C. Courcoubetis and M. Yannakakis. The complexity of probabilistic verification. *J. ACM*, 42(4):857–907, July 1995.

[6] L. de Alfaro. *Formal verification of probabilistic systems*. Ph.d. thesis, Stanford University, 1997.

[7] K. Etessami, M. Z. Kwiatkowska, M. Y. Vardi, and M. Yannakakis. Multi-objective model checking of markov decision processes. *Logical Methods in Computer Science*, 4(4), 2008.

[8] S. Even, A. L. Selman, and Y. Yacobi. The complexity of promise problems with applications to public-key cryptography. *Information and Control*, 61(2):159 – 173, 1984.

[9] E. Even-Dar, S. Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *COLT'02*, vol. 2375 of *LNCS*, pp. 255–270. Springer, 2002.

[10] O. Goldreich. On promise problems (a survey in memory of shimon even [1935-2004]). *Manuscript*, 2005.

[11] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[12] I. O. Kozine and L. V. Utkin. Interval-valued finite markov chains. *Reliable computing*, 8(2):97–113, 2002.

[13] A. Kučera and O. Stražovský. On the controller synthesis for finite-state markov decision processes. In *FSTTCS 2005*, vol. 3821 of *LNCS*, pp. 541–552. Springer, 2005.

[14] S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648, Dec. 2004.

[15] C. Ng, M. Barketau, T. Cheng, and M. 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.

[16] A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

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