The Value Problem for Multiple-Environment MDPs with Parity Objective

Krishnendu Chatterjee
IST Austria

Laurent Doyen
CNRS & LMF, ENS Paris-Saclay, France

Jean-François Raskin
Université Libre de Bruxelles, Belgium

Ocan Sankur
Université de Rennes, CNRS, Inria, France & Mitsubishi Electric R&D Centre Europe, France

Abstract

We consider multiple-environment Markov decision processes (MEMDP), which consist of a finite set of MDPs over the same state space, representing different scenarios of transition structure and probability. The value of a strategy is the probability to satisfy the objective, here a parity objective, in the worst-case scenario, and the value of an MEMDP is the supremum of the values achievable by a strategy.

We show that deciding whether the value is 1 is a PSPACE-complete problem, and even in P when the number of environments is fixed, along with new insights to the almost-sure winning problem, which is to decide if there exists a strategy with value 1. Pure strategies are sufficient for theses problems, whereas randomization is necessary in general when the value is smaller than 1. We present an algorithm to approximate the value, running in double exponential space. Our results are in contrast to the related model of partially-observable MDPs where all these problems are known to be undecidable.

2012 ACM Subject Classification Theory of computation → Logic and verification; Theory of computation → Probabilistic computation

Keywords and phrases Markov decision processes, imperfect information, randomized strategies, limit-sure winning

Digital Object Identifier 10.4230/LIPIcs...

Funding *Krishnendu Chatterjee*: ERC CoG 863818 (ForM-SMArt) and Austrian Science Fund (FWF) 10.55776/COE12;
Jean-François Raskin: PDR Weave project FORM-LEARN-POMDP funded by FNRS and DFG, and the support of the Fondation ULB;
Ocan Sankur: ANR BisoUS (ANR-22-CE48-0012) and ANR EpiRL (ANR-22-CE23-0029).

1 Introduction

We consider Markov decision processes (MDP), a well-established state-transition model for decision making in a stochastic environment. The decisions involve choosing an action from a finite set, which together with the current state determine a probability distribution over the successor state. The question of constructing a strategy that maximizes the probability to satisfy a logical specification is a classical synthesis problem with a wide range of applications [19, 11, 3, 20].

The stochastic transitions in MDPs capture the uncertainty in the effect of an action. Another form of uncertainty arises when the states are (partially) hidden to the decision-maker, as in the classical model of partially-observable MDPs (POMDP) [15, 18]. Recently,

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Figure 1 Multiple-environment MDP for the missing card (over 3-card deck). Each $M[e_i]$ represents the behavior of the MEMDP under environment e_i where card i has been removed. The environment can be identified almost-surely (with probability 1).

Figure 2 Multiple-environment MDP for the duplicate card (over 3-card deck). Each $M[e_i]$ represents the behavior of the MEMDP under environment e_i where card i has been duplicated. The environment can be identified limit-surely (with probability arbitrarily close to 1).

an alternative model of MDPs with partial information has attracted attention, the multiple-environment MDPs (MEMDP) [21], which consists of a finite set of MDPs over the same state space. Each MDP represents a possible environment, but the decision-maker does not know in which environment they are operating. The synthesis problem is then to construct a single strategy that can be executed in all environments to ensure the objective be satisfied independently of the environment. This model is natural in applications where the structure of the transitions and their probability are uncertain such as in robust planning or population models with individual variability [4, 6, 1, 23, 22].

In contrast to what previous work suggest, the two models of POMDP and MEMDP are (syntactically) incomparable: the choice of the environment in MEMDP is adversarial, which cannot be expressed in a POMDP, and the partial observability of POMDP can occur throughout the execution, whereas the uncertainty in MEMDP is only initial. In particular, MEMDP are *not* a subclass of POMDP since pure strategies are sufficient in POMDPs [17, 7] while randomization is necessary in general in MEMDPs [21, Lemma 3].

The synthesis problem has been considered for traditional ω-regular objectives, defined as parity [21] or Rabin [22] condition, in three variants: the almost-sure problem is to decide whether there exists a strategy that is winning with probability 1 in all environments, the limit-sure problem is to decide whether, for every $\varepsilon > 0$, there exists a strategy that is winning with probability at least $1 - \varepsilon$ in all environments, and the gap problem, which is an approximate version of the quantitative problem to decide, given a threshold $0 < \lambda \leq 1$, whether there exists a strategy that is winning with probability at least λ in all environments. The limit-sure problem is also called the value-1 problem, where the value of an MEMDP is defined as the supremum of the values achievable by a strategy. The value is 1 if and only if the answer to the limit-sure problem is Yes.

A classical example to illustrate the difference between almost-sure and limit-sure winning is to consider an environment consisting of 51 cards, obtained by removing one card from a standard 52-card deck (see Figure 1). The decision-maker has two possible actions: the action *sample* reveals the top card of the deck and then shuffles the cards (including the top card, which remains in the deck); the action *guess(x)*, where $1 \leq x \leq 52$ is a card, stops the game with a win if x is the missing card, and a lose otherwise. If no guess is ever made, the game is also losing. An almost-sure winning strategy is to sample until each of the 51 cards has been revealed at least once, then to make a correct guess. It is easy to see that the strategy wins with probability 1, even if there exist scenarios (though with probability 0) where some of the 51 cards are never revealed and no correct guess is made. Hence the MEMDP is almost-sure winning, and we say that it is not sure winning because a losing scenario exists in every strategy. Consider now an environment consisting of 53 cards, obtained by adding one duplicate card c to the standard deck, and the same action set and rules of the game, except that a correct guess is now the duplicate card $x = c$ (see Figure 2). The strategy that samples for a long time and then makes a guess based on the most frequent card wins with probability close to 1 – and closer to 1 as the sampling time is longer – but not equal to 1, since no matter how long is the sampling phase there is always a nonzero probability that the duplicate card does not have the highest frequency at the time of the guess. In this case, the MEMDP is limit-sure winning, but not almost-sure winning. Intuitively, the solution of almost-sure winning relies on the analysis of *revealing* transitions, which give a sure information allowing to exclude some environment (seeing card c is a guarantee that we are not in the environment where c is missing); the solution of limit-sure winning involves *learning* by sampling, which also allows to exclude some environment, but possibly with a nonzero probability to be mistaken.

For MEMDPs with two environments, it is known that the almost-sure and limit-sure problem for parity objectives are solvable in polynomial time [21, Theorem 33, Theorem 40], while the gap problem is decidable in 2-fold exponential space [21, Theorem 30] and is NP-hard, even for acyclic MEMDPs with two environments [21, Theorem 26]. With an arbitrary number of environments, the almost-sure problem becomes PSPACE-complete [22, Theorem 41], even for reachability objectives [23, Lemma 11]. For comparison, in the close model of POMDP, the decidability frontier lies between limit-sure winning and almost-sure winning: with reachability objectives, the almost-sure problem is decidable (and EXPTIME-complete [2]), whereas the limit-sure problem is undecidable [12]. The gap problem is also undecidable [16].

In this paper, we consider the limit-sure problem and the gap problem for parity objectives in MEMDPs with an arbitrary number of environments. Our main result is to show that (a) the limit-sure problem is PSPACE-complete and can be solved in polynomial time for a fixed number of environments, and (b) the gap problem can be solved in double exponential space. Correspondingly, our algorithms significantly extend the solutions that are known for two environments, relying on a non-trivial recursive (inductive) analysis.

The PSPACE upper bound is obtained by a characterization of limit-sure winning for a subclass of MEMDPs, in terms of almost-sure winning conditions (Lemma 14). A pre-processing phase transforms general MEMDPs into the subclass. We present a PSPACE algorithm to compute the pre-processing and verify the characterization. Since our algorithm relies on almost-sure winning, we also give a new characterization of almost-sure winning for parity objectives in MEMDPs (Lemma 3), which gives a conceptually simple alternative algorithm to the known solution [22]. The PSPACE lower bound straightforwardly follows from the same reduction as for almost-sure winning [22, Theorem 7]. A corollary of our characterizations is a refined strategy complexity: pure (non-randomized) strategies are sufficient for both limit-sure and almost-sure winning, which was known only for acyclic MEMDPs and almost-sure reachability objectives [23, Lemma 12], and exponential memory is sufficient. In the last part of the paper, we present an algorithm running in double exponential space for solving the gap problem, by computing an approximation of the value of the MEMDP. To win with probability at least λ in all environments, randomized strategies are more powerful [21, Lemma 3], and thus need to be considered for solving the gap problem.

In conclusion, the model of MEMDP is a valuable alternative to POMDPs, from a theoretical perspective since the limit-sure problem and gap problem are undecidable for POMDPs whereas our results establish decidability for MEMDPs, and from a practical perspective since many applications of POMDPs can be expressed by MEMDPs, as was observed previously [1, 23].

2 Definitions

A *probability distribution* on a finite set Q is a function $d : Q \to [0,1]$ such that $\sum_{q \in Q} d(q) = 1$. The support of d is $\mathsf{Supp}(d) = \{q \in Q \mid d(q) > 0\}$. A Dirac distribution assigns probability 1 to some $q \in Q$. We denote by $\mathcal{D}(Q)$ the set of all probability distributions on Q.

2.1 Markov Decision Processes

A *Markov decision process (MDP)* over a finite set A of actions is a tuple $M = \langle Q, (A_q)_{q \in Q}, \delta \rangle$ consisting of a finite set Q of *states*, a nonempty set $A_q \subseteq A$ of actions for each state $q \in Q$, and a partial probabilistic transition function $\delta : Q \times A \to \mathcal{D}(Q)$. We say that (q, a, q') is a transition if $\delta(q,a)(q') > 0$. A state $q \in Q$ is a *sink* if $\delta(q,a)(q) = 1$ for all $a \in A_q$.

A *run* of M from an initial state $q_0 \in Q$ is an infinite sequence $\pi = q_0 a_0 q_1 a_1 \dots$ of interleaved states and actions such that $a_i \in A_{q_i}$ and $\delta(q_i, a_i)(q_{i+1}) > 0$ for all $i \geq 0$. Finite prefixes $\rho = q_0 a_0 \dots q_n$ of runs ending in a state are called *histories* and we denote by $\mathsf{last}(\rho) = q_n$ the last state of ρ. We denote by $\mathsf{Hist}^\omega(M)$ (resp., $\mathsf{Hist}(M)$) the set of all runs (resp., histories) of M, and by $\mathsf{Inf}(\pi)$ the set of states that occur infinitely often along the run π.

A *sub-MDP* of M is an MDP $M' = \langle Q', (A'_q)_{q \in Q'}, \delta \rangle$ such that $Q' \subseteq Q$ and $\mathsf{Supp}(\delta(q,a)) \subseteq Q'$ for all states $q \in Q'$ and actions $a \in A'_q$ (recall the requirement that $A'_q \neq \varnothing$). Consider a set $Q' \subseteq Q$ such that for all $q \in Q'$, there exists $a \in A_q$ with $\mathsf{Supp}(\delta(q,a)) \subseteq Q'$. We define the *sub-MDP of M induced by Q'*, denoted by $M|_{Q'}$, as the sub-MDP $M' = \langle Q', (A'_q)_{q \in Q'}, \delta \rangle$ where $A'_q = \{a \in A_q \mid \mathsf{Supp}(\delta(q,a)) \subseteq Q'\}$ for all $q \in Q'$.

End-components An *end-component* of $M = \langle Q, (A_q)_{q \in Q}, \delta \rangle$ is a pair $(Q', (A'_q)_{q \in Q'})$ such that $(Q', (A'_q)_{q \in Q'}, \delta')$ is a sub-MDP of M, where δ' denotes the restriction of δ to $\{(q,a) \mid q \in Q', a \in A'_q\}$, and where the graph $\langle Q', E' \rangle$ with $E' = \{(q,q') \in Q' \times Q' \mid \exists a \in A'_q : \delta(q,a)(q') > 0\}$ is strongly connected [9, 3]. We often identify an end-component as the set $Q' \cup \{(q,a) \mid q \in Q', a \in A_q\}$ of states and state-action pairs, and we say that it is *supported* by the set Q' of states. The (componentwise) union of two end-components with nonempty intersection is an end-component, thus one can define *maximal* end-components. We denote by $\mathsf{MEC}(M)$ the set of maximal end-components of M, which is computable in polynomial time [9], and by $\mathsf{EC}(M)$ the set of all end-components of M.

Histories and Strategies A *strategy* is a function $\sigma : \mathsf{Hist}(M) \to \mathcal{D}(A)$ such that $\mathsf{Supp}(\sigma(\rho)) \subseteq A_q$ for all histories $\rho \in \mathsf{Hist}(M)$ ending in $\mathsf{last}(\rho) = q$. A strategy is *pure* if all histories are mapped to Dirac distributions. A strategy σ is *memoryless* if $\sigma(\rho) = \sigma(\rho')$ for all histories ρ, ρ' such that $\mathsf{last}(\rho) = \mathsf{last}(\rho')$. We sometimes view memoryless strategies as functions $\sigma : Q \to \mathcal{D}(A)$. A strategy σ uses *finite memory* (of size k) if there exists a right congruence $\approx$ over $\mathsf{Hist}(M)$ (i.e., such that if $\rho \approx \rho'$, then $\rho \cdot a \cdot q \approx \rho' \cdot a \cdot q$ for all $\rho, \rho' \in \mathsf{Hist}(M)$ and $(a,q) \in A \times Q$) of finite index k such that $\sigma(\rho) = \sigma(\rho')$ for all histories $\rho \approx \rho'$ with $\mathsf{last}(\rho) = \mathsf{last}(\rho')$.

Objectives An objective φ is a Borel set of runs. We denote by $\mathbb{P}^\sigma_q(M, \varphi)$ the standard probability measure on the sigma-algebra over the set of (infinite) runs of M with initial state q, generated by the cylinder sets spanned by the histories [3]. Given a history $\rho = q_0 a_0 q_1 \dots q_k$, the cylinder set $\mathrm{Cyl}(\rho) = \rho(AQ)^\omega$ has probability $\mathbb{P}^\sigma_q(M, \mathrm{Cyl}(\rho)) = \prod_{i=0}^{k-1} \sigma(q_0 a_0 q_1 \dots q_i)(a_i) \cdot \delta(q_i, a_i)(q_{i+1})$ if $q_0 = q$, and probability 0 otherwise. We say that a run ρ is compatible with strategy σ if $\mathbb{P}^\sigma_q(M, \mathrm{Cyl}(\rho)) > 0$.

We consider the following standard objectives for an MDP M:

- safety objective: given a set $T \subseteq Q$ of states, let $\mathsf{Safe}(T) = \{q_0 a_0 q_1 a_1 \ldots \in \mathsf{Hist}^\omega(M) \mid \forall i \geq 0 : q_i \in T\}$;
- reachability objective: given a set $T \subseteq Q$ of states, let $\mathsf{Reach}(T) = \{q_0 a_0 q_1 a_1 \ldots \in \mathsf{Hist}^\omega(M) \mid \exists i \geq 0 : q_i \in T\}$;
- parity objective: given a priority function $p : Q \to \mathbb{N}$, let $\mathsf{Parity}(p) = \{\pi \in \mathsf{Hist}^\omega(M) \mid \min\{p(q) \mid q \in \mathsf{Inf}(\pi)\}$ is even$\}$.

It is standard to cast safety and reachability objectives as special cases of parity objectives, using sink states. Given an objective φ, we denote by $\neg\varphi = \mathsf{Hist}^\omega(M) \setminus \varphi$ the complement of φ. We say that a run $\pi \in \mathsf{Hist}^\omega(M)$ *satisfies* φ if $\pi \in \varphi$, and that it *violates* φ otherwise.

It is known that under arbitrary strategies, with probability 1 the set $\mathsf{Inf}(\pi)$ of states occurring infinitely often along a run π is the support of an end-component [8, 9].

▶ **Lemma 1** ([8, 9])**.** *Given an MDP M, for all states $q \in Q$ and all strategies σ, we have $\mathbb{P}^\sigma_q(M, \{\pi \mid \mathsf{Inf}(\pi)$ is the support of an end-component$\}) = 1$.*

An end-component $D \in \mathsf{EC}(M)$ is *positive* under strategy σ from q if $\mathbb{P}^\sigma_q(M, \{\pi \mid \mathsf{Inf}(\pi) = D\}) > 0$. By Lemma 1, we have $\sum_{D \in \mathsf{EC}(M)} \mathbb{P}^\sigma_q(M, \{\pi \mid \mathsf{Inf}(\pi) = D\}) = 1$.

Value and qualitative satisfaction A strategy σ is winning for objective φ from q with probability (at least) α if $\mathbb{P}^\sigma_q(M, \varphi) \geq \alpha$. We denote by $\mathsf{Val}^*_q(M, \varphi) = \sup_\sigma \mathbb{P}^\sigma_q(M, \varphi)$ the *value* of objective φ from state q. A strategy σ is *optimal* if $\mathbb{P}^\sigma_q(M, \varphi) = \mathsf{Val}^*_q(M, \varphi)$.

We consider the following classical qualitative modes of winning. Given an objective φ, a state q is:

- *almost-sure winning* if there exists a strategy σ such that is winning with probability 1, that is $\mathbb{P}^\sigma_q(M, \varphi) = 1$.
- *limit-sure winning* if $\mathsf{Val}^*_q(M, \varphi) = 1$, or equivalently for all $\varepsilon > 0$ there exists a strategy σ such that $\mathbb{P}^\sigma_q(M, \varphi) \geq 1 - \varepsilon$.

We denote by $\mathsf{AS}(M, \varphi)$ and $\mathsf{LS}(M, \varphi)$ the set of almost-sure and limit-sure winning states, respectively. In MDPs, it is known that $\mathsf{AS}(M, \varphi) = \mathsf{LS}(M, \varphi)$ and pure memoryless optimal strategies exist for parity objectives φ [19, 8].

We recall that the value of a parity objective $\varphi = \mathsf{Parity}(p)$ from every state of an end-component D is the same, and is either 0 or 1, which does not depend on the precise value of the (non-zero) transition probabilities, but only on the supports $\mathsf{Supp}(\delta(q, a))$ of the transition function at the state-action pairs (q, a) in D [9]. When the value 1, there exists a pure memoryless strategy σ such that $\mathbb{P}^\sigma_q(M, \varphi) = 1$ for all states $q \in D$. If such a strategy exists, then D is said to be *φ-winning*, and otherwise *φ-losing*.

2.2 Multiple-Environment MDP

A *multiple-environment MDP (MEMDP)* over a finite set E of environments is a tuple $M = \langle Q, (A_q)_{q \in Q}, (\delta_e)_{e \in E} \rangle$, where $M[e] = \langle Q, (A_q)_{q \in Q}, \delta_e \rangle$ is an MDP that models the behaviour of the system in the environment $e \in E$. The state space is identical in all $M[e]$ ($e \in E$), only the transition probabilities may differ. We sometimes refer to the environments of M as the MDPs $\{M[e] \mid e \in E\}$ rather than the set E itself. For $E' \subset E$, let $M[E']$ be the MEMDP M over set E' of environments. We denote by $M[\neg e]$ the MEMDP M over environments $E \setminus \{e\}$, and by $\cup_{e \in E} M[e]$ the MDP $\langle Q, (A_q)_{q \in Q}, \delta_\cup \rangle$ such that $\delta_\cup(q, a)$ is the uniform distribution over $\bigcup_{e \in E} \mathsf{Supp}(\delta_e(q, a))$ for all $q \in Q$ and $a \in A$.

A transition $t = (q, a, q')$ is *revealing* in M if $K_t = \{e \in E \mid q' \in \mathsf{Supp}(\delta_e(q, a))\}$ is a strict subset of E ($K_t \subsetneq E$). We say that K_t, which is the set of environments where the transition $t = (q, a, q')$ is possible, is the *knowledge* after observing transition t. An MEMDP is in *revealed form* if for all revealing transitions $t = (q, a, q')$, the state q' is a sink in all environments, that is $\mathsf{Supp}(\delta_e(q', a)) = \{q'\}$ for all environments $e \in E$ and all actions $a \in A_{q'}$. By extension, we call knowledge after a history ρ the set of environments in which all transitions of ρ are possible.

Decision Problems We are interested in synthesizing a *single* strategy σ with guarantees in *all* environments, without knowing in which environment σ is executing. We consider reachability, safety, and parity objectives.

A state q is *almost-sure winning* in M for objective φ if there exists a strategy σ such that in all environments $e \in E$, we have $\mathbb{P}^\sigma_q(M[e], \varphi) = 1$, and we call such a strategy σ almost-sure winning. A state q is *limit-sure winning* in M for objective φ if for all $\varepsilon > 0$, there exists a strategy σ such that in all environments $e \in E$ we have $\mathbb{P}^\sigma_q(M[e], \varphi) \geq 1 - \varepsilon$, and we say that such a strategy σ is $(1 - \varepsilon)$-winning.

We denote by $\mathsf{AS}(M, \varphi)$ (resp., $\mathsf{LS}(M, \varphi)$) the set of all almost-sure (resp., limit-sure) winning states in M for objective φ. We consider the *membership problem for almost-sure (resp., limit-sure) winning*, which asks whether a given state q is almost-sure (resp., limit-sure) winning in M for objective φ. We refer to these membership problems as *qualitative* problems.

We are also interested in the *quantitative* problems. Given MEMDP M, a parity objective φ, and probability threshold $\alpha \geq 0$, we are interested in the existence of a strategy σ satisfying $\mathbb{P}^\sigma_q(M[e], \varphi) \geq \alpha$ for all $e \in E$. We present an approximation algorithm for the quantitative problem, solving the *gap problem* consisting, given MEMDP M, state q, parity objective φ, and thresholds $0 < \alpha < 1$ and $\varepsilon > 0$, in answering

- Yes if there exists a strategy σ such that for all $e \in E$, we have $\mathbb{P}^\sigma_q(M[e], \varphi) \geq \alpha$,
- No if for all strategies σ, there exists $e \in E$ with $\mathbb{P}^\sigma_q(M[e], \varphi) < \alpha - \varepsilon$,
- and arbitrarily otherwise.

The 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 [10, 13].

Results We solve the membership problem for limit-sure winning with parity objectives φ (i.e., deciding whether a given state q is limit-sure winning, that is $q \in \mathsf{LS}(M, \varphi)$), providing a PSPACE algorithm with a matching complexity lower bound, and showing that the problem is solvable in polynomial time when the number of environments is fixed. Our solution relies on the solution of almost-sure winning, which is known to be PSPACE-complete for reachability [23] and Rabin objectives [22]. We revisit the solution of almost-sure winning and give a simple characterization for safety objectives (which is also PSPACE-complete), that can easily be extended to parity objectives. A corollary of our characterization is that pure (non-randomized) strategies are sufficient for both limit-sure and almost-sure winning, which was known only for acyclic MEMDPs and reachability objectives [23, Lemma 12].

For the gap problem, we present an double exponential-space procedure to approximate the value α that can be achieved in all environments, up to an arbitrary precision ε.

3 Almost-Sure Winning

It is known that the membership problem for almost-sure winning in MEMDPs is PSPACE-complete with reachability objectives [23] as well as with Rabin objectives [22], an expressively equivalent of the parity objectives. We revisit the membership problem for almost-sure winning with parity and safety objectives, as it will be instrumental to the solution of limit-sure winning. We present a conceptually simple characterization of the winning region for almost-sure winning, from which we derive a PSPACE algorithm, thus matching the known complexity for almost-sure Rabin objectives. A corollary of our characterization is that pure (non-randomized) strategies are sufficient for both limit-sure and almost-sure winning, which was known only for acyclic MEMDPs and reachability objectives [23, Lemma 12].

▶ **Theorem 2** ([23],[22]). *The membership problem for almost-sure winning in MEMDPs with a reachability, safety, or Rabin objective is PSPACE-complete.*

To solve the membership problem for a safety or parity objective φ, we first convert M into an MEMDP M' in revealed form with state space $Q \uplus \{q_{\mathsf{win}}, q_{\mathsf{lose}}\}$ and each revealing transition $t = (q, a, q')$ in M is redirected in M' to q_{win} if $q' \in \mathsf{AS}(M[K_t], \varphi)$ is almost-sure winning when the set of environments is the knowledge K_t after observing transition t, and to q_{lose} otherwise. In order to decide if $q' \in \mathsf{AS}(M[K_t], \varphi)$, we need to solve the membership problem for an MEMDP with strictly fewer environments than in M as $K_t \subsetneq E$, which will lead to a recursive algorithm. The base case of the solution is MEMDPs with one environment, which is equivalent to plain MDPs.

It is easy to see that $\mathsf{AS}(M, \varphi) \cup \{q_{\mathsf{win}}\} = \mathsf{AS}(M', \varphi \cup \mathsf{Reach}(q_{\mathsf{win}}))$ for all prefix-independent objectives φ, and we can transform the objective $\varphi \cup \mathsf{Reach}(q_{\mathsf{win}})$ into an objective of the same type as φ (for example, if φ is a parity objective then assigning the smallest even priority to q_{win} turns the objective $\varphi \cup \mathsf{Reach}(q_{\mathsf{win}})$ into a pure parity objective).

Hence, the main difficulty is to solve the membership problem for MEMDP in revealed form.

3.1 Safety

Although safety objectives are subsumed by parity objectives which we solve in the next section, we give here a simpler algorithm specifically for safety, and also prove PSPACE-hardness in this case.

The safety objective has the property that almost-sure winning is equivalent to sure winning, where a strategy is sure winning if all runs compatible with the strategy satisfy the objective. Intuitively, if some runs does not satisfy the safety objective $\mathsf{Safe}(T)$, then it contains a state outside T after a finite prefix, thus with positive probability (the probability of the finite prefix). In the sure-winning mode, we can consider the probabilistic choices to be adversarial, which entails that only the support of the probability distributions in the transition function is relevant.

It follows that, as long as the knowledge remains the set E of all environments a winning strategy for a safety objective can play all actions that are safe (i.e., that ensure the successor state remains in the winning region) in all environments. We obtain the following property: almost-sure winning for a safety objective in a MEMDP M in revealed form is equivalent to almost-sure winning in the MDP $\cup_{e \in E} M[e]$.

An algorithm for solving almost-sure safety is as follows: (1) for each revealing transition $t = (q, a, q')$ in M, decide if $q' \in \mathsf{AS}(M[K_t], \mathsf{Safe}(T))$ (using a recursive call), and redirect the transition t to q_{win} or q_{lose} accordingly, transforming M into revealed form; (2) assuming M is in revealed form, compute the almost-sure winning states $W = \mathsf{AS}(M_{\cup}, \mathsf{Safe}(T))$ where $M_{\cup} = \cup_{e \in E} M[e]$ is an MDP. Return $W \setminus \{q_{\mathsf{win}}\}$. The depth of recursive calls is bounded by the number of environments, and the almost-sure safety in MDPs can be solved in polynomial time, namely, in time $O(|Q|^2|A|)$. It follows that almost-sure safety in MEMDPs can be solved in PSPACE, and in time $O(|Q|^2 \cdot |A| \cdot 2^{|E|})$. A PSPACE lower bound can be established by a similar reduction from QBF as for reachability, the constructed MEMDP being acyclic [23].

Note that for a fixed number of environments, the membership problem for almost-sure safety in MEMDPs is solvable in polynomial time by our algorithm since the depth of the recursion is then constant. This is also the case in Theorem 2 as shown in [23].

3.2 Parity

By definition, the almost-sure winning region $W = \mathsf{AS}(M, \mathsf{Parity}(p))$ for a parity objective in an MEMDP M is such that there exists a strategy σ that is almost-sure winning for the parity objective from every state $q \in W$ in every MDP $M[e]$ (where e is an environment of M). In contrast, we show the following characterization (note the order of the quantifiers).

▶ **Lemma 3.** *Given an MEMDP M in revealed form with state space Q, if $W \subseteq Q$ is such that in every environment e, from every state $q \in W$, there exists a strategy σ_e that is almost-sure winning for the parity objective* $\mathsf{Parity}(p)$ *in $M|_W[e]$ from q, then $W \subseteq \mathsf{AS}(M, \mathsf{Parity}(p))$. Moreover, for all $q \in W$, there exists a pure $(|Q| \cdot |E|)$-memory strategy ensuring* $\mathsf{Parity}(p)$ *from q in M.*

Proof. For each environment $M[e]$, consider a memoryless strategy σ_e almost-surely winning for the objective $\mathsf{Parity}(p)$ in $M|_W[e]$ from every state of W. Recall that almost-sure winning strategies can be assumed to be memoryless in MDPs with single environments; and that one can build a single memoryless strategy that is almost-surely winning from all winning states. Let $\mathsf{EC}(\sigma_e) = \{D \in \mathsf{EC}(M[e]) \mid \exists q \in W : \mathbb{P}_q^{\sigma_e}(M[e], \mathsf{Inf} = D) > 0\}$ be the set of positive end-components under strategy σ_e. Note that the least priority in an end-component $D \in \mathsf{EC}(\sigma_e)$ is even since the parity objective is satisfied with probability 1.

Let $E = \{1, \dots, k\}$ be the set of environments of M. We construct a pure almost-sure winning strategy σ for the MEMDP M as follows, where initially $e = 1$:

(1) play according to σ_e for $|W|$ steps;

(2) if the current state is q_{win} or belongs to a positive end-component $D \in \mathsf{EC}(\sigma_e)$, keep playing according to σ_e forever. Otherwise, increment e (modulo k) and go to (1).

The strategy σ uses memory of size at most $|Q| \cdot |E|$ since $W \subseteq Q$.

Fix environment $f \in E$. We show that strategy σ is almost-sure winning in $M[f]$. Because all strategies σ_e are defined in $M|_W$, the region W is never left while playing σ, and during phase (1) of the strategy there is a lower-bounded probability to reach an end-component $D \in \mathsf{EC}(\sigma_e)$ when $e = f$.

We show that eventually phase (2) is executed forever with probability 1, that is, some end-component $D \in \mathsf{EC}(\sigma_e)$ for some e is reached with probability 1. Towards contradiction, assume that phase (1) of the strategy σ is executed infinitely often with positive probability p. Then phase (1) for $e = f$ and σ_f is also executed infinitely often and it follows that, conditioned on phase (1) being executed infinitely often, a positive end-component $D \in \mathsf{EC}(\sigma_f)$ is reached with probability 1; hence phase (2) is executed forever from that point on. Thus with probability $1 - p + p = 1$ phase (1) is executed only finitely often, contradicting our assumption.

As phase (2) of the strategy σ is eventually executed forever with probability 1, let e be the corresponding environment (i.e., such that σ plays according to σ_e) and let $D \neq \{q_{\mathsf{win}}\}$ be the reached end-component of $M[e]$ (the other case where q_{win} is reached is trivial). If some transition of D is not present in f, then it must be a revealing transition in e, thus leading in $M[e]$ to q_{win} outside D, which is impossible since D is an end-component in $M[e]$. Hence all transitions of D are present in all environments.

Algorithm 1 AS_Parity(M, p)

```
Input  : $M = \langle Q, (A_q)_{q \in Q}, (\delta_e)_{e \in E} \rangle$ an MEMDP, $p : Q \to \mathbb{N}$ a priority function.
Output: The winning region $\mathsf{AS}(M, \mathsf{Parity}(p))$ for almost-sure parity.
begin
                                                        /* pre-processing */
1      $M' \leftarrow M$
2      add two sink states $q_{\mathsf{win}}, q_{\mathsf{lose}}$ to $M'$
3      define $p(q_{\mathsf{win}}) = 0$ and $p(q_{\mathsf{lose}}) = 1$
4      foreach revealing transition $t = (q, a, q')$ in $M$ do
                                                        /* $K_t \subsetneq E$ */
5          if $q' \in \mathsf{AS\_Parity}(M[K_t], p)$ then
6              replace $t$ by $(q, a, q_{\mathsf{win}})$ in $M'$
           else
7              replace $t$ by $(q, a, q_{\mathsf{lose}})$ in $M'$
8      $M \leftarrow M'$
                                                        /* $M$ is in revealed form */
9      $P \leftarrow \varnothing$; $P' \leftarrow \varnothing$
10     repeat
11         $P \leftarrow \cap_{e \in E} \mathsf{AS}(M[e], \mathsf{Parity}(p))$        /* $M[e]$ and $\cup_{e \in E} M[e]$
12         $P' \leftarrow \mathsf{AS}(\cup_{e \in E} M[e], \mathsf{Safe}(P))$            are MDPs */
13         $M \leftarrow M|_{P'}$
       until $P'$ is unchanged
14     return $P' \setminus \{q_{\mathsf{win}}\}$
end
```

We show that σ is almost-sure winning in f. The result is immediate if D is an end-component of $M[f]$ (in particular if $f = e$). If D is not an end-component of $M[f]$, then in $M[f]$ the strategy would leave D and reach q_{win}, thus σ is almost-sure winning as well in that case. ◀

The characterization in the first part of Lemma 3 holds simply because parity objectives are prefix-independent (runs that differ by a finite prefix are either both winning or both losing), and thus the characterization holds for all prefix-independent objectives.

The converse of Lemma 3 is immediate, which entails that the almost-sure winning region $W = \mathsf{AS}(M, \mathsf{Parity}(p))$ is the largest set of states satisfying the condition in Lemma 3. We exploit this characterization in Algorithm 1 to compute the winning region for almost-sure parity. After transforming the MEMDP into revealed form (through recursive calls to the algorithm), we compute the winning region for almost-sure parity in each environment (line 11), and then the set P' of states from which we can remain in the intersection P of all these winning regions (line 12). We iterate this process on $M|_{P'}$ until a fixpoint $P = P'$ is reached.

It is easy to see that the fixpoint satisfies the characterization of Lemma 3, and thus $P' \subseteq \mathsf{AS}(M, \mathsf{Parity}(p)) \cup \{q_{\mathsf{win}}\}$. Also by the proof of Lemma 3, we can construct a pure almost-sure winning $(|Q| \cdot |E|)$-memory strategy from all states in P', and define (recursively, for each subset of the environments) a pure almost-sure winning strategy from the states that were replaced by q_{win} in the revealed form, with a total memory size at most $|Q| \cdot |E| \cdot 2^{|E|}$, corresponding to the memory bound from Lemma 3 for each subset $K \subseteq E$ of environments (representing the belief, i.e., the set of environments where the current history is possible).

To show the converse inclusion, we show the invariant that every state $q \in Q \setminus P'$ is not almost-sure winning in M: for all strategies σ from q, in some environment $M[e]$ the set P is left with positive probability (along some history ρ). Given a state $q' \in Q \setminus P$ reached in $M[e]$, there is an environment $f \in E$ where the parity objective is violated with positive probability under σ from q'. The crux is to show that the state q' is reached with positive probability in $M[f]$ as well. Towards contradiction, assume that the history ρ from q to q' (in $M[e]$) is not possible in $M[f]$. Then ρ contains a revealing transition in $M[e]$, and $q' = q_{\mathsf{win}} \in P$, which is a contradiction since $q' \in Q \setminus P$. Hence, in $M[f]$ with strategy σ the parity objective is violated with positive probability.

Algorithm 1 can be implemented in PSPACE by a similar argument as for almost-sure safety: the depth of recursive calls is bounded by the number of environments, both almost-sure safety and almost-sure parity can be solved in polynomial time in MDPs, and the repeat-loop runs at most $|Q|$ times. The algorithm runs in polynomial time if the number of environments is fixed. The PSPACE-hardness follows from Theorem 2.

▶ **Theorem 4.** *The membership problem for almost-sure parity in MEMDPs is PSPACE-complete. Pure exponential-memory strategies are sufficient for almost-sure winning in MEMDPs with parity (thus also reachability and safety) objectives. When the number of environments is fixed, the problem is solvable in polynomial time.*

The time complexity of Algorithm 1 is established as follows. Each recursive call, corresponds to a subset of the initial environment set E that we can compute once and tabulate. In each call, the second loop runs at most $|Q|$ times, and the set of almost-sure winning states for parity conditions (that is, the set $\mathsf{AS}(M[e], \mathsf{Parity}(p))$) can be computed in time $O(|Q| \cdot |\delta|)$ [3]. Since $|\delta|$ is in $O(|Q|^2 \cdot |A|)$, each recursive call takes $O(|Q|^4 \cdot |E| \cdot |A|)$ time, and overall, this is $O(|Q|^4 \cdot |E| \cdot |A| \cdot 2^{|E|})$.

Note that pure exponential-memory strategies for almost-sure parity in MEMDPs are provided by Lemma 3. The algorithm for almost-sure parity can be used to solve almost-sure safety with optimal PSPACE complexity, although the specific algorithm for safety is slightly simpler (the repeat-loop can be replaced by just line 12 where $P = T$ is the set of states defining the safety objective $\mathsf{Safe}(T)$).

The PSPACE procedure can be implemented in exponential time by solving all subproblems and storing their solutions. Moreover, for large numbers of environments, the exponent in the complexity can be made to depend only on the size of M. In fact, intuitively, two environments with identical supports yield the same result so one can derive a dynamic programming solution where at most one environment per support is solved.

Define the *support* of a probabilistic transition relation $\delta : Q \times A \to \mathcal{D}(Q)$ as the family of supports of its transitions, that is, $\mathsf{Supp}(\delta) = (\mathsf{Supp}(\delta(q, a)))_{(q,a) \in Q \times A}$. Define the support of a family of transition relations as $\mathsf{Supp}((\delta_e)_{e \in E}) = \{\mathsf{Supp}(\delta_e) \mid e \in E\}$.

Two environments δ_e and δ_f are said to be *equivalent* if they have the same support. One can check whether two environments are equivalent in polynomial time, by going through all triples (q, a, q') and verifying that $\delta_e(q, a, q') = 0$ iff $\delta_f(q, a, q') = 0$.

Almost sure parity in MEMDPs does not depend on the precise probability values in the given environments in M but only on their supports.

In addition to Theorem 4, we can obtain a complexity bound whose exponent is independent of the number of environments (Theorem 6), using the following result: if in two environments, the support of the transition relation is the same, we can discard one of the environment (all strategies that are almost-sure winning in one are also almost-sure winning in the other one, as shown in Lemma 5) and thus consider at most one environment for each support. Here, we denote by $\mathsf{Supp}((\delta_e)_{e\in E}) = (\mathsf{Supp}(\delta_e))_{e\in E}$ where $\mathsf{Supp}(\delta_e)$ denotes the set of transitions with positive probability under δ_e.

▶ **Lemma 5.** *Consider two MEMDPs $M_i = \langle Q, (A_q)_{q\in Q}, (\delta_e)_{e\in E_i}\rangle$ for $i = 1, 2$, with the same state and action sets, and with the same supports of their transition relation, $\mathsf{Supp}((\delta_e)_{e\in E_1}) = \mathsf{Supp}((\delta_e)_{e\in E_2})$. Given a parity condition $\mathsf{Parity}(p)$, for all states q and all finite-memory strategies σ, the following equivalence holds: $\mathbb{P}^\sigma_q[M_1[e], \mathsf{Parity}(p)] = 1$ for all $e \in E_1$ if and only if $\mathbb{P}^\sigma_q[M_2[e], \mathsf{Parity}(p)] = 1$ for all $e \in E_2$. In particular, $\mathsf{AS}(M_1, \mathsf{Parity}(p)) = \mathsf{AS}(M_2, \mathsf{Parity}(p))$.*

Proof. Given state q and finite-memory strategy σ, assume that $\mathbb{P}^\sigma_q[M_1[e_1], \mathsf{Parity}(p)] = 1$ for all $e_1 \in E_1$. Consider any $e_2 \in E_2$, and let $e_1 \in E_1$ be such that $\mathsf{Supp}(\delta_{e_1}) = \mathsf{Supp}(\delta_{e_2})$; such a e_2 exists by the hypothesis $\mathsf{Supp}((\delta_e)_{e\in E_1}) = \mathsf{Supp}((\delta_e)_{e\in E_2})$. Consider the Markov chain obtained as the product of the MDP $M_1[e_1]$ with the Moore machine describing the finite-memory strategy σ. Because $\mathbb{P}^\sigma_q[M_1[e_1], \mathsf{Parity}(p)] = 1$, all bottom strongly connected components (BSCC) in this product are winning for $\mathsf{Parity}(p)$ (i.e., the smallest priority of their states is even). But the product of $M_2[e_2]$ and the Moore machine for σ have the same set of BSCCs since the supports are identical. It follows that $\mathbb{P}^\sigma_q[M_2[e_2], \mathsf{Parity}(p)] = 1$. By symmetry, this proves the first statement.

It follows that $\mathsf{AS}(M_1, \mathsf{Parity}(p)) = \mathsf{AS}(M_2, \mathsf{Parity}(p))$ since finite-memory strategies suffice for almost-sure parity in MEMDPs by Theorem 4.

◀

▶ **Theorem 6.** *The membership problem for almost-sure parity for an MEMDP $M = \langle Q, (A_q)_{q\in Q}, (\delta_e)_{e\in E}\rangle$ can be solved in time $O((|E|^2 + |Q|^4 \cdot |E| \cdot |A|) \cdot 2^{\min(|E|, 2^{|Q|^2 \cdot |A|})})$.*

Proof. Consider an MEMDP $M = \langle Q, (A_q)_{q\in Q}, (\delta_e)_{e\in E}\rangle$ and a parity objective φ. If $|E| \leq 2^{|Q|^2|A|}$, then we apply the PSPACE procedure from Theorem 4. The number of recursive calls is then bounded by $2^{|E|}$, and each call itself takes polynomial time, so the result follows.

Otherwise, we scan the set of environments given as input, and store a subset E' of these: we include an environment e in E' if and only if none of the previously stored environments is equivalent to e. This takes $O(|E|^2)$ time. This yields a subset with at most $2^{|Q|^2|A|}$ environments, with at most one representative for each possible support. We then apply the recursive algorithm on the MEMDP $M[E']$, which yields the same result as if it was applied to $M = M[E]$ by Lemma 5. ◀

4 Limit-Sure Winning

We refer to the examples of the duplicate card and the missing card in Section 1 to illustrate the difference between limit-sure and almost-sure winning. We present in Section 4.1 two other scenarios where limit-sure winning and almost-sure winning do not coincide, which will be useful to illustrate the key ideas in the algorithmic solution.

4.1 Examples

In the example of Figure 3, the set $D = \{q_1, q_2\}$ is an end-component in both environments e_1 and e_2 (the actions are shown in the figures only when relevant, that is in q_2). However,

Figure 3 An end-component $\{q_1, q_2\}$ with different transition probabilities in environments e_1 and e_2.

Figure 4 The set $\{q_1, q_2\}$ is an end-component in e_2, not in e_1.

the transition probabilities from q_1 are different in the two environments e_1 and e_2, and intuitively we can learn (with high probability) in which environment we are by playing c for a long enough (but finite) time and collecting the frequency of the visits to q_1 and q_2. Then, in order to reach the target q_3, if there are more q_1's than q_2's in the history we play a in q_2, otherwise b. The intuition is that the histories with more q_1's than q_2's have a high probability (more than $1 - \varepsilon$) in $M[e_1]$ and a small probability (less than ε) in $M[e_2]$, where ε can be made arbitrarily small (however not 0) by playing c for sufficiently long. Hence q_1 is limit-sure winning, but not almost-sure winning.

In the second scenario (Figure 4), the transition probabilities do not matter. The objective is to visit some state in $\{q_3, q_4, q_5\}$ infinitely often (those states have priority 0, the other states have priority 1). The state q_1 is limit-sure winning, but not almost-sure winning. To win with probability $1 - \varepsilon$, a strategy can play a (in q_2) for a sufficiently long time, then switch to playing b (unless q_5 was reached before that). The crux is that playing a does not harm, as it does not leave the limit-sure winning region, but ensures in at least one environment (namely, e_1) that the objective is satisfied with probability 1 (by reaching q_5). This allows to "discard" the environment e_1 if q_5 was not reached, and to switch to a strategy that is winning with probability at least $1 - \varepsilon$ in e_2, namely by playing b. With an arbitrary number of environments, the difficulty is to determine in which order the environments can be "discarded".

Note that the transition (q_2, a, q_1) is not revealing, since it is present in both environments. However, after crossing this transition a large number of times, we can still learn that the environment is e_2 (and be mistaken with arbitrarily small probability). In contrast, the transition (q_2, a, q_5) is revealing and the environment is e_1 with certainty upon crossing that transition.

To solve the membership problem for limit-sure parity, we first convert M into a revealed-form MEMDP M', similar to the case of almost-sure winning, with the obvious difference that revealing transitions $t = (q, a, q')$ of $M[e]$ are redirected in $M'[e]$ to q_{win} if $q' \in \mathsf{LS}(M[K_t], \varphi)$ is limit-sure winning when the set of environments is the knowledge K_t after observing transition t. Thus, we aim for a recursive algorithm, where the base case is limit-sure winning in MEMDPs with one environment, which are equivalent to plain MDPs, for which limit-sure and almost-sure parity coincide. Note that the examples of Figure 3 and Figure 4 are in revealed form.

4.2 Common End-Components and Learning

A *common end-component (CEC)* of an MEMDP $M = \langle Q, (A_q)_{q \in Q}, (\delta_e)_{e \in E} \rangle$ is a pair (Q', A') that is an end-component in $M[e]$ for all environments $e \in E$. A CEC D is *trivial* if it contains a single state. D is said *winning* for a parity condition $\mathsf{Parity}(p)$, if for all $e \in E$, there is a strategy in $M[e]$ which, when started inside D, ensures $\mathsf{Parity}(p)$ with probability 1. Notice that since D is a common end-component, such a strategy ensures $\mathsf{Parity}(p)$ with probability 1 in $M[e]$ iff it does in $M[e']$.

We note that the common end-components of an MEMDP are the end-components of the MDP $\cup_{e\in E}M[e]$ assuming M is in revealed form, and thus can be computed using standard algorithm for end-components [9].

▶ **Lemma 7.** *Consider an MEMDP M in revealed form. The common end-components of M are exactly the end-components of $\cup_{e\in E}M[e]$.*

Proof. Consider a common end-component D of M. Because in each $M[e]$, all state-action pairs in D stay inside D, and D is strongly connected, this is also the case in $\cup_{e\in E}M[e]$; thus D is an end-component of the latter.

Conversely, consider an end-component D of $\cup_{e\in E}M[e]$. If D consists of a single sink state, then it is indeed a common end-component. Otherwise D contains more than one state. We show that all state-action pairs (q, a) of D must have the same support in all environments, and it follows that D is an end-component in every environment, thus a common end-component. By contradiction, if a transition (q, a, q') with $(q, a) \in D$ exists in $M[e]$ but not in $M[f]$, then it is revealing and q' is a sink state. Hence D is not strongly connected in $\cup_{e\in E}M[e]$ because D does not consist of a single sink state. ◀

A CEC may have different transition probabilities in different environments. We call a CEC *distinguishing* if it contains a transition (q, a, q') (called a distinguishing transition) such that $\delta_e(q, a)(q') \neq \delta_f(q, a)(q')$ for some environments $e, f \in E$. Given a distinguishing transition (q, a, q') and environment e, define $K_1 = \{f \in E \mid \delta_f(q, a)(q') = \delta_e(q, a)(q')\}$ and $K_2 = E \setminus K_1$. We say that (K_1, K_2) is a *distinguishing partition* of D that is *induced* by the distinguishing transition (q, a, q') and environment e.

Distinguishing transitions can be used to learn the partition (K_1, K_2), that is to guess (correctly with high probability) whether the current environment is in K_1 or K_2, as in the example of Figure 3, where the set $D = \{q_1, q_2\}$ is a distinguishing end-component with distinguishing transition $(q_1, \cdot, q_2)$ and partition $(\{e_1\}, \{e_2\})$. A distinguishing CEC may have several distinguishing transitions and induced partitions.

We formalize how a strategy can distinguish between K_1 and K_2 with high probability inside a distinguishing CEC. First let us recall Hoeffding's inequality.

▶ **Theorem 8** (Hoeffding's Inequality [14]). *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identical Bernoulli variables with $\mathbb{P}[X_i] = p$, and write $S_n = X_1 + \ldots + X_n$. For all $t > 0$, $\mathbb{P}[S_n - \mathbb{E}[S_n] \geq t] \leq e^{-2t^2/n}$, and $\mathbb{P}[\mathbb{E}[S_n] - S_n \geq t] \leq e^{-2t^2/n}$.*

Given a distinguishing CEC with distinguishing partition (K_1, K_2) induced by a transition (q, a, q'), a strategy can sample the distribution $\delta_e(q, a)$ by repeating the following two phases: first, use a pure memoryless strategy to almost-surely visit q, then play action a; by repeating this long enough (the precise bound depends on a given ε and is derived from Theorem 8) while storing the frequency of visits to q' in the second phase, we can learn and guess in which block K_i belongs the environment, with sufficiently small probability of mistake to ensure winning with probability $1 - \varepsilon$.

▶ **Lemma 9.** *Given an MEMDP M containing a distinguishing common end-component D with partition (K_1, K_2) induced by a distinguishing transition, and parity objective φ, for all states q_0 in D, all pairs of strategies σ_1, σ_2, and all $\varepsilon > 0$, there exists a strategy σ such that:*

$$\mathbb{P}^{\sigma}_{q_0}(M[e], \varphi) \geq (1 - \varepsilon)\mathbb{P}^{\sigma_1}_{q_0}(M[e], \varphi) \text{ for all } e \in K_1,$$
$$\mathbb{P}^{\sigma}_{q_0}(M[e], \varphi) \geq (1 - \varepsilon)\mathbb{P}^{\sigma_2}_{q_0}(M[e], \varphi) \text{ for all } e \in K_2.$$

Moreover, the strategy σ is pure if both σ_i are pure; and if each strategy σ_i uses a memory of size m_i, then σ uses finite memory of size $m_1 + m_2 + \lceil 8\frac{\log(1/\varepsilon)}{\eta^2}^2\rceil$ where $\eta = \min\left(\{|\delta_e(q,a)(q') - \delta_f(q,a)(q')| \mid e,f \in E, q,q' \in Q, a \in A\} \setminus \{0\}\right)$.

Proof. Consider $M = \langle Q, (A_q)_{q\in Q}, (\delta_e)_{e\in E}\rangle$ and $D = (Q', (A'_q)_{q\in Q'})$ as in the statement of the lemma, and let $q_0 \in D$.

Consider a distinguishing transition (q,a,q') and environment e_0 that induces the distinguishing partition (K_1, K_2). Consider $\varepsilon > 0$, and define $N = \lceil\frac{2\log(1/\varepsilon)}{\eta^2}\rceil$,

The strategy σ runs in two phases. In the first phase, the goal is to estimate the distribution of (q,a,q'). For this, it executes a pure memoryless strategy which has a nonzero probability of reaching q while staying in D (such a strategy can be defined based on the supports of state-action pairs of D) and keeps two counters: $c_{q,a}$ that counts the number of times the state-action pair (q,a) is selected; and $c_{q,a,q'}$ the number of times the transition (q,a,q') is observed. The second round of the strategy starts when $c_{q,a} = N$. Note that this happens with probability 1. Then, we go back to q_0 (with probability 1), and we switch to

- σ_1 if $\left|\frac{c_{q,a,q'}}{c_{q,a}} - \delta_{e_0}(q,a)(q')\right| < \eta/2$,
- σ_2 otherwise.

We now analyze this strategy and show that because N is sufficiently large, the estimation error is bounded, so that we obtain the desired result.

In each environment e, at each visit at q and choice of a, we have a Bernoulli trial with mean $\delta_e(q,a)(q')$, and $c_{q,a,q'}$ is the number of successful trials. By Hoeffding's inequality (Theorem 8), we have

$$\mathbb{P}^{\sigma}_{q_0}\left(M[e], |c_{q,a,q'}/c_{q,a} - \delta_e(q,a)(q')| \geq \eta/2 \;\middle|\; c_{q,a} = N\right) \leq e^{-2N(\frac{\eta}{2})^2} \leq \varepsilon.$$

Thus, in $M[e]$ with $e \in K_i$, the probability of not switching to σ_i is at most ε. It follows that $\mathbb{P}^{\sigma}_{q_0}(M[e],\varphi) \geq (1-\varepsilon)\mathbb{P}^{\sigma_i}_{q_0}(M[e],\varphi)$.

The memory requirement comes from the fact that σ must store two counters up to N values, and it has two modes (before and after reaching $c_{q,a} = N$). ◀

It follows that the membership problem for limit-sure winning can be decomposed into subproblems where the set of environments is one of the blocks K_i in the partition.

▶ **Lemma 10.** *Given an MEMDP M containing a distinguishing common end-component D with a partition (K_1, K_2) induced by a distinguishing transition, and a parity objective φ the following equivalence holds: $D \subseteq \mathsf{LS}(M,\varphi)$ if and only if $D \subseteq \mathsf{LS}(M[K_1],\varphi)$ and $D \subseteq \mathsf{LS}(M[K_2],\varphi)$.*

Proof. Immediate consequence of Lemma 9. ◀

4.3 Characterization and Algorithm

Here, we assume that MEMDPs are in revealed form with sink states q_{win} and q_{lose}.

We show that the winning region $W = \mathsf{LS}(M,\varphi)$ for limit-sure parity is a closed set: from every state $q \in W$, there exists an action a ensuring in all environments that all successors of q are in W. We call such actions *limit-sure safe* for q. We show in Lemma 11 that a limit-sure safe action always exists in limit-sure winning states. Note that playing actions that are *not* limit-sure safe may be useful for limit-sure winning, as in the example of Figure 3 where action a is limit-sure safe, but action b is not (from q_2).

By definition of limit-sure winning, if a state q is not limit-sure winning, there exists $\varepsilon_q > 0$ such that for all strategies σ, there exists an environment $e \in E$ such that $\mathbb{P}_q^\sigma(M[e], \varphi) < 1 - \varepsilon_q$. We denote by $\varepsilon_0 = \min\{\varepsilon_q \mid q \in Q \setminus \mathsf{LS}(M, \varphi)\}$ a uniform bound.

▶ **Lemma 11.** *Given an MEMDP M (in revealed form) over environments E, a parity objective φ, and a state q, if $q \in \mathsf{LS}(M, \varphi)$ is limit-sure winning, then there exists an action a such that for all environments $e \in E$, all successors of q are limit-sure winning, i.e* $\mathsf{Supp}(\delta_e(q, a)) \subseteq \mathsf{LS}(M, \varphi)$.

Proof. Consider $q \in \mathsf{LS}(M, \varphi)$ and let $0 < \varepsilon < \frac{\nu\varepsilon_0}{|A|}$, where A is the set of actions in M, and ν is a lower bound on the smallest nonzero transition probability (in all environments), and ε_0 is the uniform bound defined above. Let σ be a strategy ensuring φ from q with probability at least $1 - \varepsilon$ in all environments.

Towards contradiction, assume that there is no limit-sure safe action from state q. Let a be the action chosen by σ with the highest probability at the history q, that is $a = \arg\max_a \sigma(q)(a)$, and thus $\sigma(q)(a) \geq \frac{1}{|A|}$. By our assumption, there exists an environment $e \in E$ and a state $t \notin \mathsf{LS}(M, \varphi)$ (in particular $t \neq q_{\mathsf{win}}$) such that $\delta_e(q, a)(t) > 0$, hence $\delta_e(q, a)(t) \geq \nu$. It is immediate that $t \neq q_{\mathsf{lose}}$ as otherwise the strategy σ would ensure φ with probability at most $1 - \nu \leq 1 - \varepsilon$ from q. So $t \notin \{q_{\mathsf{win}}, q_{\mathsf{lose}}\}$ and therefore $\delta_e(q, a)(t) \geq \nu$ in all environments e. By definition of the uniform bound ε_0, there exists an environment e such that $\mathbb{P}_t^\sigma(M[e], \varphi) \leq 1 - \varepsilon_0$, hence from q we have $\mathbb{P}_q^\sigma(M[e], \neg\varphi) \geq \frac{\nu\varepsilon_0}{|A|} > \varepsilon$, in contradiction to σ ensuring φ with probability at least $1 - \varepsilon$ from q. We conclude that there exists a limit-sure safe action from q. ◀

Given an MEMDP M, consider the limit-sure winning region $W = \mathsf{LS}(M, \varphi)$ for $\varphi = \mathsf{Parity}(p)$. For the purpose of the analysis, consider the (memoryless) randomized strategy σ_{LS} that plays uniformly at random all limit-sure safe actions in every state $q \in W$, which is well-defined by Lemma 11.

Consider an arbitrary environment e, and an end-component D in $M[e]$ that is positive under σ_{LS} (recall Lemma 1 and the definition afterward). There are three possibilities:

1. D is not a common end-component (as in the example of Figure 4, for $D = \{q_1, q_2\}$ in $M[e_2]$), that is, D is not an end-component in some environment e' (in the example $e' = e_1$), then we can learn (and be mistaken with arbitrarily small probability) that we are not in e', reducing the problem to an MEMDP with fewer environments (namely, $M[\neg e']$);
2. D is a common end-component and is distinguishing (as in the example of Figure 3, for $D = \{q_1, q_2\}$), then we can also learn a distinguishing partition (K_1, K_2) and reduce the problem to MEMDPs with fewer environments (namely, $M[K_1]$ and $M[K_2]$);
3. D is a common end-component and is non-distinguishing, then we show in Lemma 12 below that D is almost-sure winning ($D \subseteq \mathsf{AS}(M, \varphi)$), obviously in all environments.

▶ **Lemma 12.** *Given an MEMDP M over environments E (in revealed form), a parity objective φ, and a state q, if $q \in \mathsf{LS}(M, \varphi)$, then all non-distinguishing common end-components D that are positive under strategy σ_{LS} from q in $M[e]$ (for some $e \in E$) are almost-sure winning for φ (that is $D \subseteq \mathsf{AS}(M, \varphi)$).*

Proof. Consider a positive non-distinguishing common end-component D as in the statement of the lemma. Using Lemma 11, note that $D \subseteq \mathsf{LS}(M, \varphi)$ since σ_{LS} plays only limit-sure safe actions and D is a common end-component.

Assume towards contradiction that D is not almost-sure winning for the parity objective φ. It follows that in M, all strategies that play only limit-sure safe actions ensure the parity objective φ with probability 0 from all states in D (in all environments since D is a common end-component).

Denote by Ω_{safe} the set of all runs that contain only limit-sure safe actions. For all strategies σ (in M), and $q \in D$ we have $\mathbb{P}^{\sigma}_{q}(M[e], \varphi \mid \Omega_{safe}) = 0$ (for all $e \in E$) and therefore:

$$\begin{aligned}\mathbb{P}^{\sigma}_{q}(M[e], \varphi) &= \mathbb{P}^{\sigma}_{q}(M[e], \varphi \mid \Omega_{safe}) \cdot \mathbb{P}^{\sigma}_{q}(M[e], \Omega_{safe}) \\ &\quad + \mathbb{P}^{\sigma}_{q}(M[e], \varphi \mid \neg\Omega_{safe}) \cdot \mathbb{P}^{\sigma}_{q}(M[e], \neg\Omega_{safe}) \\ &= \mathbb{P}^{\sigma}_{q}(M[e], \varphi \mid \neg\Omega_{safe}) \cdot \mathbb{P}^{\sigma}_{q}(M[e], \neg\Omega_{safe}) \\ &\leq 1 - \mathbb{P}^{\sigma}_{q}(M[e], \neg\varphi \mid \neg\Omega_{safe})\end{aligned}$$

Given $\varepsilon < \frac{\varepsilon_0 \cdot \nu}{|E|}$ where ν is the smallest positive probability in M, we show that there exists an environment $e \in E$ such that $\mathbb{P}^{\sigma}_{q}(M[e], \varphi) < 1 - \varepsilon$, which entails that q is not limit-sure winning for φ, establishing a contradiction since $q \in D \subseteq \mathsf{LS}(M, \varphi)$. It will follow that D is almost-sure winning for φ and conclude the proof.

By definition of limit-sure safe actions, to every pair (q, a) such that $a \in A_q$ is not limit-sure safe in q, we can associate an environment e such that:

$$\mathsf{Supp}(\delta_e(q, a)) \cap (Q \setminus \mathsf{LS}(M, \varphi)) \neq \varnothing,$$

and thus from some state $q' \in \mathsf{Supp}(\delta_e(q, a))$, we have $\mathbb{P}^{\sigma}_{q'}(M[e], \varphi) \leq 1 - \varepsilon_0$ where ε_0 is the uniform bound for non-limit-sure winning states. Assuming that a non-limit-sure safe action is played by σ, since there are finitely many environments, by the pigeonhole principle there is an environment e such that with probability at least $\frac{1}{|E|}$ an action that is not limit-sure safe and associated with e is played, which leads with probability at least ν to a state outside $\mathsf{LS}(M, \varphi)$. It follows that $\mathbb{P}^{\sigma}_{q}(M[e], \neg\varphi \mid \neg\Omega_{safe}) \geq \varepsilon_0 \cdot \frac{\nu}{|E|} > \varepsilon$ and thus $\mathbb{P}^{\sigma}_{q}(M[e], \varphi) < 1 - \varepsilon$, which concludes the proof. ◀

Our approach to compute the limit-sure winning states is to first identify the distinguishing CECs that are limit-sure winning. We can compute the maximal CECs using Lemma 7, and note that a maximal CEC containing a distinguishing CEC is itself distinguishing, so it is sufficient to consider maximal CECs. By Lemma 10, we can decide if a given distinguishing CEC is limit-sure winning using a recursive procedure on MEMDPs with fewer environments. We show in Lemma 13 below that we can replace the limit-sure CECs by a sink state q_{win}.

▶ **Lemma 13.** *Given an MEMDP M with parity objective φ and a set $T \subseteq \mathsf{LS}(M, \varphi)$ of limit-sure winning states, we have $\mathsf{LS}(M, \varphi) = \mathsf{LS}(M, \varphi \cup \mathsf{Reach}(T))$.*

Proof. The inclusion $\mathsf{LS}(M, \varphi) \subseteq \mathsf{LS}(M, \varphi \cup \mathsf{Reach}(T))$ is immediate since $\varphi \subseteq \varphi \cup \mathsf{Reach}(T)$.

To show the converse inclusion, consider $q \in \mathsf{LS}(M, \varphi \cup \mathsf{Reach}(T))$ and show that $q \in \mathsf{LS}(M, \varphi)$. Given $\varepsilon > 0$, let $\varepsilon_1 = \frac{\varepsilon}{2}$ and let σ be a strategy such that $\mathbb{P}^{\sigma}_{q}(M, \varphi \cup \mathsf{Reach}(T)) \geq 1 - \varepsilon_1$. We construct a strategy τ that satisfies the objective φ with probability at least $1 - \varepsilon$ as follows: for all histories ρ, if ρ does not visit T, then let $\tau(\rho) = \sigma(\rho)$; otherwise, consider the suffix ρ' of ρ after the first visit to a state $t \in T$, and let σ_t be strategy that ensures φ is satisfied with probability at least $1 - \varepsilon_1$ from t (such a strategy exists since $T \subseteq \mathsf{LS}(M, \varphi)$). Define $\tau(\rho) = \sigma_t(\rho')$. We easily show below that $\mathbb{P}^{\tau}_{q}(M, \varphi) \geq 1 - \varepsilon$, establishing that $q \in \mathsf{LS}(M, \varphi)$:

$$\begin{aligned}
\mathbb{P}^{\tau}_{q}(M, \varphi) &= \mathbb{P}^{\tau}_{q}(M, \varphi \cap \mathsf{Reach}(T)) + \mathbb{P}^{\tau}_{q}(M, \varphi \cap \neg\mathsf{Reach}(T)) \\
&= \mathbb{P}^{\tau}_{q}(M, \varphi \mid \mathsf{Reach}(T)) \cdot \mathbb{P}^{\tau}_{q}(M, \mathsf{Reach}(T)) + \mathbb{P}^{\tau}_{q}(M, \varphi \cap \neg\mathsf{Reach}(T)) \\
&= \mathbb{P}^{\tau}_{q}(M, \varphi \mid \mathsf{Reach}(T)) \cdot \mathbb{P}^{\sigma}_{q}(M, \mathsf{Reach}(T)) + \mathbb{P}^{\sigma}_{q}(M, \varphi \cap \neg\mathsf{Reach}(T)) \\
&\qquad\qquad\qquad\text{(since } \tau \text{ agrees with } \sigma \text{ as long as } T \text{ is not reached)} \\
&\geq (1 - \varepsilon_1) \cdot \mathbb{P}^{\sigma}_{q}(M, \mathsf{Reach}(T)) + \mathbb{P}^{\sigma}_{q}(M, \varphi \cap \neg\mathsf{Reach}(T)) \\
&\geq (1 - \varepsilon_1) \cdot \mathbb{P}^{\sigma}_{q}(M, \mathsf{Reach}(T)) + (1 - \varepsilon_1) \cdot \mathbb{P}^{\sigma}_{q}(M, \varphi \cap \neg\mathsf{Reach}(T)) \\
&\geq (1 - \varepsilon_1) \cdot \mathbb{P}^{\sigma}_{q}(M, \varphi \cup \mathsf{Reach}(T)) \geq (1 - \varepsilon_1)^2 \geq 1 - \varepsilon.
\end{aligned}$$

◀

We can now assume that MEMDPs contain no limit-sure winning distinguishing CEC, and present a characterization for the remaining possibility, illustrated by the scenario of Figure 4, where playing the action a (in q_2, forever) ensures, in some environment (namely, e_1), almost-sure satisfaction of the parity objective while remaining inside the limit-sure winning region in all other environments.

▶ **Lemma 14.** *Consider an MEMDP M (in revealed form) over environments E with $|E| \geq 2$, that contains no limit-sure winning distinguishing common end-component, and a parity objective φ. Writing $T_e = \mathsf{LS}(M[\neg e], \varphi)$, we have the following:* $\mathsf{LS}(M, \varphi) = \mathsf{AS}\Big(M, \mathsf{Reach}\Big(\bigcup_{e \in E} \mathsf{AS}(M[e], \varphi \cap \mathsf{Safe}(T_e))\Big)\Big)$.

Proof. First we show the inclusion

$$\mathsf{LS}(M, \varphi) \subseteq \mathsf{AS}(M, \mathsf{Reach}(\textstyle\bigcup_{e \in E} \mathsf{AS}(M[e], \varphi \cap \mathsf{Safe}(T_e)))).$$

Consider the (memoryless) strategy σ_{LS} that plays all limit-sure safe actions uniformly at random from every state in $\mathsf{LS}(M, \varphi)$. The strategy σ_{LS} is well-defined by Lemma 11 and to establish the inclusion, we show that, from every state $q \in \mathsf{LS}(M, \varphi)$, it is almost-sure winning (in all environments $e' \in E$) for the objective $\mathsf{Reach}(\bigcup_{e \in E} \mathsf{AS}(M[e], \varphi \cap \mathsf{Safe}(T_e)))$.

Consider an arbitrary environment $e' \in E$ and an arbitrary end-component D that is positive under σ_{LS} in $M[e']$. Since positive end-components are reached with probability 1 (Lemma 1), it is sufficient to show that for all such D, there exists an environment $e \in E$ such that every state in D is almost-sure winning for the objective $\varphi \cap \mathsf{Safe}(T_e)$ in $M[e]$. We consider two cases:

- if D is a common end-component, then we show that D is non-distinguishing. Note that D must be limit-sure winning, by definition of limit-sure safe actions (played by σ_{LS}). It follows by the assumption of the lemma that D is non-distinguishing and therefore almost-sure winning for φ (in all environments) by Lemma 12. We take $e = e'$ and it is easy to see that there exists an almost-sure winning strategy for φ from D (that stays in D), which is also almost-sure winning for $\varphi \cap \mathsf{Safe}(T_e)$.
- otherwise D is not a common end-component, and there exists an environment e where D is not an end-component. We first show that all transitions of D are present in $M[e]$, since otherwise D would contain a revealing transition, thus leading to a state that is a sink in all environments (revealed form). Then D being strongly connected would not contain another state, and thus in particular all transitions in D would be present in $M[e]$.

It follows that playing σ_{LS} from D in $M[e]$ ensures with probability 1 that a (revealing) transition not present in $M[e']$ is executed, which leads to q_{win} since σ_{LS} never leaves the limit-sure winning region (by definition of limit-sure safe actions). Hence φ is satisfied with probability 1 in $M[e]$ while playing only limit-sure safe actions, thus remaining in the limit-sure winning region $\mathsf{LS}(M, \varphi) \subseteq \mathsf{LS}(M[\neg e], \varphi) = T_e$, thereby satisfying $\mathsf{Safe}(T_e)$ as well. This shows that in $M[e]$, the states in D are almost-sure winning for the objective $\varphi \cap \mathsf{Safe}(T_e)$.

For the converse inclusion, given a state q and a pure[1] strategy σ that is almost-sure winning for objective $\mathsf{Reach}(\bigcup_{e\in E} \mathsf{AS}(M[e], \varphi \cap \mathsf{Safe}(T_e)))$ (in all environments), we show that for all $\varepsilon > 0$ there is a pure strategy τ that ensures that φ is satisfied with probability at least $1 - \varepsilon$ (from q in all environments).

Given $\varepsilon > 0$, let τ be the strategy that plays as follows:

(1) play like σ until a state $t \in \bigcup_{e\in E} \mathsf{AS}(M[e], \varphi \cap \mathsf{Safe}(T_e))$ is reached, and let $e \in E$ be an environment such that from t there is a (pure memoryless) strategy σ_t that is almost-sure winning in $M[e]$ for the objective $\varphi \cap \mathsf{Safe}(T_e)$;

(2) play like σ_t for $k \cdot |Q|$ steps, where k is such that $(1 - \nu^{|Q|})^k \leq \varepsilon$ (where ν is the smallest positive probability in M);

(3) if the current state belongs to a positive end-component D_t of σ_t (in $M[e]$), then keep playing like σ_t (forever); otherwise switch to a strategy that ensures that φ is satisfied with probability at least $1 - \varepsilon$ from the current state in all environments of $E \setminus \{e\}$ – such a strategy exists because from t the strategy σ_t ensures the objective $\mathsf{Safe}(T_e)$ is satisfied almost-surely (and thus surely as well).

Consider an arbitrary environment $e \in E$, and show that $\mathbb{P}^{\tau}_{q}(M[e], \varphi) \geq 1 - \varepsilon$, which establishes that q is limit-sure wining, $q \in \mathsf{LS}(M, \varphi)$.

First note that phase (2) (and thus also phase (3)) is reached with probability 1, and let e_t be the environment corresponding to the state t reached at the end of phase (1). We consider two cases:

- if $e_t = e$, then by standard analysis the probability that after phase (2) a positive end-component of σ_t is *not yet* reached is at most $(1 - \nu^{|Q|})^k \leq \varepsilon$ since within $|Q|$ steps a positive end-component is reached with probability at least $\nu^{|Q|}$. Hence with probability at least $1 - \varepsilon$, a positive (winning since σ_t almost-sure winning in $M[e]$ for the objective φ) end-component of σ_t is reached and the strategy σ_t is played forever in phase (3), thus winning with probability at least $1 - \varepsilon$.
- otherwise $e_t \neq e$ and we consider the following cases in phase (3):

  (a) if the strategy σ_t is played forever, then either the set D_t (which is an end-component in $M[e_t]$) is never left, or it is left (via a revealing transition, as D_t is not left in $M[e_t]$) and since σ_t ensures $\mathsf{Safe}(T_{e_t})$ the sink q_{win} is reached in $M[e]$, thus in both cases the objective φ is satisfied (with probability 1);

  (b) otherwise, by construction the strategy τ switches to a strategy that ensures φ is satisfied with probability at least $1 - \varepsilon$.

In all cases, the objective φ holds with probability at least $1-\varepsilon$, showing that $\mathbb{P}^{\tau}_{q}(M[e], \varphi) \geq 1 - \varepsilon$ as claimed.

◀

Algorithm 2 $\mathsf{LS\_Parity}(M, p)$

```
Input  : $M = \langle Q, (A_q)_{q \in Q}, (\delta_e)_{e \in E} \rangle$ an MEMDP, $p : Q \to \mathbb{N}$ a priority function.
Output: The winning region $\mathsf{LS}(M, \mathsf{Parity}(p))$ for limit-sure parity.
begin
    if $|E| = 1$ then return $\mathsf{AS}(M, \mathsf{Parity}(p))$
                                                        /* pre-processing */
    put $M$ in revealed form (defined in Section 3)
    $\mathsf{MEC} \leftarrow$ maximal end-components of the MDP $\cup_{e \in E} M'[e]$
    for $D \in \mathsf{MEC}$ do
        if $D$ is distinguishing in $M$ then
            Let $(K_1, K_2)$ be a distinguishing partition in $D$
            if $D \subseteq \mathsf{LS\_Parity}(M[K_1], p) \cap \mathsf{LS\_Parity}(M[K_2], p)$ then
                replace $D$ by sink $q_{\mathsf{win}}$ in $M$ with $p(q_{\mathsf{win}}) = 0$
                                    /* $M$ is in revealed form and Lemma 14 applies */
    for $e \in E$ do
        $T_e = \mathsf{LS\_Parity}(M[\neg e], p)$
    $Q \leftarrow \mathsf{AS}\Big(M, \mathsf{Reach}\Big(\bigcup_{e \in E} \mathsf{AS}(M[e], \mathsf{Parity}(p) \cap \mathsf{Safe}(T_e))\Big)\Big)$
    return $Q \setminus \{q_{\mathsf{win}}\}$
end
```

Algorithm Overview Given a MEMDP $M = (Q, (A_q)_{q \in Q}, (\delta_e)_{e \in E})$, the algorithm proceeds by recursion on the size of the environment set E (Algorithm 2). The base case is that of a singleton set E where $\mathsf{LS}(M, \varphi) = \mathsf{AS}(M, \varphi)$ and this can be computed in polynomial time.

Assume $|E| \geq 2$. We first convert M into an MEMDP M' in revealed form with state space $Q \uplus \{q_{\mathsf{win}}, q_{\mathsf{lose}}\}$ and each revealing transition $t = (q, a, q')$ in M is redirected in M' to q_{win} if $q' \in \mathsf{LS}(M[K_t], \varphi)$ is limit-sure winning when the set of environments is the knowledge K_t after observing transition t, and to q_{lose} otherwise. Notice that each query $q' \in \mathsf{LS}(M[K_t], \varphi)$ uses a set K_t that is strictly smaller than E.

We now assume that M is in revealed form and we compute the maximal end-components of the MDP $\cup_{e \in E} M[e]$; these are maximal common end-components of M by Lemma 7. For each distinguishing maximal CEC D, we determine whether it is limit-sure winning using the condition of Lemma 10, namely that $D \subseteq \mathsf{LS}(M[K_i], \varphi)$ (for $i = 1, 2$) where (K_1, K_2) is a partition of E induced by a distinguishing transition of D, which is computed by a recursive calls to the algorithm. We replace D by q_{win} if it is limit-sure winning, which yields an MEMDP without limit-sure winning distinguishing CECs, and we can apply Lemma 14: for each environment $e \in E$, we compute $T_e = \mathsf{LS}(M[\neg e], \varphi)$ which is done by $|E|$ separate recursive calls, and we compute the sets $\mathsf{AS}(M[e], \varphi \cap \mathsf{Safe}(T_e))$ using standard MDP algorithms (we restrict the state space to T_e and compute the almost-sure winning states for φ). We then solve the almost-sure reachability problem in M for the target set $\bigcup_{e \in E} \mathsf{AS}(M[e], \varphi \cap \mathsf{Safe}(T_e))$.

[1] By Theorem 4, pure strategies are sufficient for almost-sure winning in MEMDPs.

Thus each recursive step takes polynomial time (besides the recursive calls), and because each recursive call decreases the size of E, the depth of the recursion is bounded by $|E|$. It follows that the procedure runs in polynomial space. The PSPACE lower bound follows from the same reduction as for almost-sure winning [22, Theorem 7], since the MEMDP constructed in the reduction is acyclic, thus almost-sure and limit-sure winning coincide.

Note that Lemma 14 constructs a pure strategy that achieves the objective with probability at least $1 - \varepsilon$ from the limit-sure winning states, and that the strategies constructed in Lemmas 10 and 13 to witness limit-sure winning are also pure (in Lemma 10, the construction assumes that pure strategies are sufficient for fewer environments, which allows a proof by induction since pure strategies are sufficient in MDPs, *i.e.* in a single environment).

▶ **Theorem 15.** *The membership problem for limit-sure parity objectives in MEMDPs is PSPACE-complete and pure exponential-memory strategies are sufficient, i.e., if a state q is limit-sure winning, then for all $\varepsilon > 0$ there exists a pure exponential-memory strategy that ensures the objective is satisfied with probability at least $1 - \varepsilon$ from q. When the number of environments is fixed, the problem is solvable in polynomial time.*

The time complexity of Algorithm 2 is established as follows. Let us consider a single recursive call. The maximal end-components of $\cup_{e \in E} M'[e]$ can be computed in $O(|Q| \cdot |\delta|)$ where $|\delta|$ denotes the number of transitions. Then, determining whether each MEC is distinguishing, and replacing them with sink states can be done in time $O(|\delta| \cdot |E|)$ since one needs to go over each transition and check whether their probability differs in two environments. The last step requires solving almost-sure parity and safety for MDPs defined for each $e \in E$, which can be done in time $O(|E| \cdot |Q| \cdot |\delta|)$ (similarly as in the discussion following Theorem 4). The most costly operation is almost-sure reachability for the MEMDP M, which by Theorem 4 takes $O(|Q|^4 \cdot |E| \cdot |A| \cdot 2^{|E|})$. There are $2^{|E|}$ recursive calls (the algorithm can be run once for each subset of E using memoization), so overall we get $O(|Q|^4 \cdot |E| \cdot |A| \cdot 2^{2|E|})$.

We do not know if a technique similar to that of Theorem 6 can be used for the limit-sure case to obtain an exponent independent of $|E|$.

5 The Gap Problem

The goal of this section is to give a procedure that solves the gap problem for parity objectives. For this, we show that an arbitrary strategy in M can be imitated by a finite-memory one (with a computable bound on the memory size) while achieving the same probability of winning up to ε in all environments. Once this is established, we show how to guess a finite-memory strategy of the appropriate size in order to solve the gap problem.

To establish the memory bound for such an ε-approximation, we need a few intermediate lemmas. First, we define a transformation on MEMDPs consisting in collapsing non-distinguishing maximal CECs (MCECs) of the MEMDP M; the resulting MEMDP is denoted $\mathsf{purge}(M)$. We show that M and $\mathsf{purge}(M)$ have the same probabilities of satisfaction of the considered parity objective under all environments.

Intuitively, removing non-distinguishing MCECs ensures that in $\mathsf{purge}(M)$, under all strategies, with high probability, within a fixed number of steps, either a maximal CEC is reached (which is either distinguishing, or non-distinguishing but trivial – recall that a trivial CEC contains a single absorbing state.) or enough samples are gathered to improve the knowledge about the current environment, as shown in Section 5.2 This observation will help us constructing the finite-memory strategy inductively since in each case the knowledge can be improved correctly with high probability: in trivial MCECs, the strategy is extended arbitrarily; inside distinguishing MCECs, the strategy can be extended so that it stays inside the MCEC while sampling distinguishing transitions with any desired precision as in Lemma 9; last, if no MCECs are reached but enough samples are gathered along the way, we prove that the knowledge can also be improved with high probability. The final strategy is obtained by combining finite-memory strategies constructed inductively for smaller sets of environments. This is done in Section 5.3

Maximal Common End-Components Revisited

We extend the definition of common end-components (CEC) which, in Section 4.2, were defined assuming MEMDPs are in revealed form. In this section, MEMDPs are **not** assumed to be in revealed form: in fact, upon observing a revealing transition, we cannot conclude recursively since we cannot determine which value vector must be achieved in the recursive call. Here, we define a CEC for MEMDP $M = \langle Q, A, (\delta_e)_{e \in E} \rangle$ as a pair (Q', A') such that for all $e \in E$, $\langle Q', A', \delta_e \rangle$ is an end-component of $M[e]$. A maximal CEC (MCEC) is a CEC which does not contain a smaller CEC.

There are two types of MCECs:

- MCEC (Q', A') is non-distinguishing if for all $q \in Q'$, and $a \in A'(q)$, the distributions $\delta_e(q, a)$ and $\delta_{e'}(q, a)$ are identical for all $e, e' \in E$;
- MCEC (Q', A') is distinguishing otherwise.

While non-distinguishing MCECs have state-action pairs with identical supports in all environments, a distinguishing MCEC may contain revealing transitions, that is, state-action pairs (q, a) with different supports in different environments. This is the difference with Section 4. The only result we need from Section 4.2 is Lemma 9 which holds for the new definition of distinguishing MCECs: in fact, we do require that (Q', A') is an end-component (i.e., closed and strongly connected) in all environments, so revealing transitions are simply seen as distinguishing transitions, and thanks to the strong connectivity of (Q', A') in all environments, one can define a strategy that samples a distinguishing transition a desired number of times.

As previously, a MCEC D is *trivial* if it contains a single state.

In terms of computability, we cannot use Lemma 7 to compute MCECs since this is only valid for MEMDPs in revealed form. The ε-gap procedure given in this section does not actually compute MCECs; these are only used in the proof of the existence of a finite-memory strategy (Lemma 24). Nevertheless, for completeness, let us describe how MCECs can be computed in polynomial time. For $|E| = 1$, the MCECs are exactly the maximal end-components (MECs) of $M[e]$ where $E = \{e\}$. For $|E| \geq 2$, we pick an environment $e \in E$, and compute the MECs of $M[e]$. For each MEC D of $M[e]$, we recursively compute the MCECs of D in the MEMDP $M[E \setminus \{e\}]$. This is sound because a MCEC, being an end-component in all environments, is necessarily a subset of some MEC in each $M[e]$; so by restricting the search for MCECs to MECs of some $M[e]$, we do not discard any MCECs. Furthermore, each recursive call splits the state space to disjoint sets, so we get an overall polynomial-time complexity.

Given an MEMDP M over environments E, the notation $\mathbb{P}^\sigma_q(M, \varphi)$ refers to the vector of probability values $(\mathbb{P}^\sigma_q(M[e], \varphi))_{e \in E}$.

Figure 5 An MEMDP M with two environments (left) and the construction $\mathsf{purge}(M)$ (right). Transition probabilities are uniform. Here D is the MCEC defined by the pairs $\{(q_3, a), (q_4, a)\}$, and D' is the MCEC defined by $\{(q_5, a), (q_6, a)\}$. The priority function is omitted, we assume that D is winning (e.g., by assigning priority 0 to q_3 and q_4) and that D' is losing (e.g., by assigning priority 1 to q_5 and q_6).

5.1 Purge: Removing Non-Distinguishing MCECs

We first describe a transformation that collapses non-distinguishing MCECs, and keeps only trivial ones. Since all trivial MCECs can be classified into winning and losing for the objective φ, we assume that the only non-distinguishing MCECs in the resulting MEMDP are called q_{win} and q_{lose}. The intuition is that non-distinguishing MCECs are not useful to refine information in order to distinguish environments, so when a strategy visits such a MCEC, one can assume that it will either stay inside forever (either if the MCEC is φ-winning, or if there is no outgoing transition), or leave it as soon as possible (if the MCEC is φ-losing).

Observe that a distinguishing MCEC can contain a smaller non-distinguishing CEC. The transformation described here only collapses MCECs that are non-distinguishing, and not those smaller non-distinguishing CECs that are contained in MCECs.

Given an MEMDP $M = \langle Q, A, (\delta_e)_{e \in E} \rangle$, define the MEMDP $\mathsf{purge}(M) = \langle Q', A', (\delta'_e)_{e \in E} \rangle$ where Q' contains all states of Q except those that belong to non-distinguishing MCECs; and for each non-distinguishing MCEC D, we add a fresh state s_D to Q', and redirect all transitions that enter a state of D in M to s_D in M'. We define the map $f : Q \to Q'$ by mapping all states of non-distinguishing MCECs D to s_D, and as the identity for other states.

We add a fresh action stay which from s_D goes to a winning absorbing state q_{win} if D is φ-winning, and to a losing absorbing state q_{lose} otherwise. For each pair $(q, a) \in D$ such that $\mathsf{Supp}(\delta(q, a))$ is not included in D, we add a fresh action $F_{(q,a)}$ from s_D with $\delta'_e(s_D, F_{(q,a)})(q') = \sum_{q'' \in f^{-1}(q')} \delta_e(q, a, q'')$ for all $e \in E$. (These state-action pairs can leave D in some environments, so F stands for the *frontier* of D.)

Given the set of MCECs, $\mathsf{purge}(M)$ can be computed in polynomial time. However, the ε-gap procedure we give does not actually compute $\mathsf{purge}(M)$; this construction is only used for proving the existence of a finite-memory strategy of bounded memory size.

▶ **Example 16.** An example of this construction is given in Fig. 5 for MEMDP M with two environments e_1, e_2. Here $\{(q_2, b)\}$ is an end-component in $M[e_2]$ but not in $M[e_1]$ due to the edge to q_3 so this is not a CEC, and is not collapsed in $\mathsf{purge}(M)$. The MCEC D defined by $\{(q_3, a), (q_4, a)\}$ has a single frontier action $F_{(q_4,b)}$. In $M[e_1]$, we have $\delta'_{e_1}(s_D, F_{(q_4,b)}, s_{D'}) = 2/3$ since $\delta_{e_1}(q_4, b, q5) + \delta_{e_1}(q_4, b, q6) = 2/3$ (since the probabilities are uniform), and $\delta'_{e_1}(s_D, F_{(q_4,b)}, s_D) = 1/3$. In $M[e_2]$, the latter edge is missing, so $\delta'_{e_2}(s_D, F_{(q_4,b)}, s_{D'}) = 1$.

▶ **Lemma 17.** *For all MEMDPs M, the only non-distinguishing MCECs of* $\mathsf{purge}(M)$ *are the trivial* q_{win} *and* q_{lose}.

Proof. Let $D = (Q', A')$ be any non-distinguishing MCEC in M'. D must contain a state of the form $s_{D'}$ since otherwise this is also a MCEC of M, and the construction would have collapsed it. We consider the component in M given by the inverse image of D by f. Formally, let $Q'' = f^{-1}(Q') \subseteq Q$, and for each $q'' \in Q''$, define $A''(q'') = \{a \in A(q'') \mid \forall e \in E : \mathsf{Supp}(\delta_e(q, a)) \subseteq Q''\}$.

Then for each state of the form $q_{D'}$ in D, (Q'', A'') contains all state-action pairs of D'. But D is strongly connected in each $M'[e]$, and all non-distinguishing MCECs D' of M that were collapsed are also strongly connected in each $M[e]$ by definition, (Q'', A'') is also strongly connected in each $M[e]$, thus a MCEC in M.

Now, (Q'', A'') cannot be distinguishing, since the construction only collapses MCECs, so no subset of (Q'', A'') can be collapsed in M'; and (Q'', A'') would remain untouched and be distinguishing in M' as well. So (Q'', A'') is non-distinguishing; but in this case, it is collapsed into a trivial MCEC in M', so D is trivial. ◀

To relate the histories of M to those of $\mathsf{purge}(M)$, we introduce the function $h \mapsto \mathsf{purge}(h)$ which, intuitively, maps the state of a non-distinguishing MCECs D to the state s_D, removes the state-actions pairs that stay in D, and replaces the state-action pairs (q, a) having a transition that leaves D by a new action $F_{(q,a)}$. Formally, $\mathsf{purge}(h)$ is obtained from $h = q_1 a_1 \ldots q_n$ by applying the following transformation: for each non-distinguishing MCEC $D = (Q', A')$ of M,

1. Replace the maximal suffix of h of the form $q_i a_i \ldots q_n$ such that for all $i \leq k \leq n$, $q_k \in Q'$ and $a_k \in A'(q_k)$, if such a suffix exists, by s_D;
2. Remove all maximal factors of h of the form $q_i a_i \ldots q_j a_j$ satisfying $q_k \in Q'$ and $a_k \in A'(q_k)$ for all $i \leq k \leq j$;
3. Replace each pair $q_i a_i$ with $q_i \in Q'$ and $a_i \notin A'(q_i)$ by $s_D F_{(q_i,a_i)}$;

▶ **Example 18.** In the MEMDP of Fig. 5, with D containing the pairs (q_3, a) and (q_4, a), for $h = q_1 a q_3 a q_4 a q_3 a q_4 b q_4 b q_5$, we get $\mathsf{purge}(h) = q_1 a s_D F_{q_4,b} s_D F_{q_4,b} q_5$. Here we first apply rule 2 above to the factor $q_3 a q_4 a q_3 a$, and get $q_1 a q_4 b q_4 b q_5$; then an application of rule 3 yields $\mathsf{purge}(h) = q_1 a s_D F_{q_4,b} s_D F_{q_4,b} q_5$. For the history $h' = q_1 a q_3 a q_4 a q_3 a q_4$, we would get by rule 1, $\mathsf{purge}(h') = q_1 a s_D$.

We establish a relation between M and $\mathsf{purge}(M)$ in Lemmas 19 and 20. These will be used to give a memory bound for strategies for the quantitative case in Lemma 24.

In the Lemma 19, we only establish an inequality. This is because a given strategy σ of M may not be optimal within a non-distinguishing MCEC, while the construction $\mathsf{purge}(M)$ is based on the assumption that optimal strategies are used within each MCECs.

▶ **Lemma 19.** *Consider an MEMDP $M = \langle Q, A, (\delta_e)_{e \in E} \rangle$, and objective $\varphi = \mathsf{Parity}(p)$, and the map $f : Q \to Q'$ relating states of M and those of $\mathsf{purge}(M) = \langle Q', A', (\delta'_e)_{e \in E} \rangle$. For all $q \in Q$, and strategy σ for M, there exists σ' with $\mathbb{P}^{\sigma}_{q}(M, \varphi) \leq \mathbb{P}^{\sigma'}_{f(q)}(\mathsf{purge}(M), \varphi)$.*

Proof. Let us write $M' = \mathsf{purge}(M)$. Consider $q \in Q$, and a strategy σ for M. We define σ' for M' as follows. For all histories h of M', and action $a \in A'(\mathsf{last}(h))$, we define

$$\sigma'(h)(a) = \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(ha) \mid \mathsf{purge}^{-1}(h)\right]$$

for some arbitrary $e \in E$ for which $\mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h)\right] > 0$, if such $e \in E$ exists; and otherwise define $\sigma'(h)$ arbitrarily. This quantity does not depend on e since, assuming $\mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h)\right] > 0$,

$$\begin{aligned}
&\mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(ha) \mid \mathsf{purge}^{-1}(h)\right] \\
&= \sum_{\rho \in \mathsf{purge}^{-1}(h)} \mathbb{P}^{\sigma}_{q}\left[M[e], \rho a' \mid \rho\right] \mathbb{P}^{\sigma}_{q}\left[M[e], \rho \mid \mathsf{purge}^{-1}(h)\right] \\
&= \sum_{\rho \in \mathsf{purge}^{-1}(h)} \sigma(\rho)(a') \mathbb{P}^{\sigma}_{q}\left[M[e], \rho \mid \mathsf{purge}^{-1}(h)\right],
\end{aligned}$$

where $a' = b$ if a has the form $F_{(\_,b)}$, and $a' = a$ otherwise (in which case we have $a \in A(\mathsf{last}(\rho))$). Moreover, $\mathbb{P}^{\sigma}_{q}[M[e], \rho \mid \mathsf{purge}^{-1}(h)]$ does not depend on e here since $\mathsf{purge}^{-1}(h)$ determines the outcomes of all transitions whose probability distributions differ among environments because these were not erased by $\mathsf{purge}(\cdot)$, and these probability distributions are identical in the remaining transitions since they belong to non-distinguishing MCECs.

For a history h of M' that ends in a state of the form s_D, and with $\mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h)\right] > 0$, we let σ' take the action stay with probability $\mathbb{P}^{\sigma}_{q}[M[e], \mathsf{purge}^{-1}(h)D^{\omega} \mid \mathsf{purge}^{-1}(h)]$, where D^{ω} denotes the set of all runs that stay inside D. This probability is similarly independent from the particular choice of e.

We prove that for all histories h of M' that do not contain stay, $a \in A'(\mathsf{last}(h)) \setminus \{\mathsf{stay}\}$, and $e \in E$,

$$\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h\right] = \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h)\right], \tag{1}$$

$$\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h \cdot a\right] = \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h \cdot a)\right], \tag{2}$$

$$\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h \cdot \mathsf{stay}\right] = \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h)D^{\omega}\right]. \tag{3}$$

We proceed by induction on the length of h to prove the above three properties.

Initially, if $|h| = 1$, then $h = f(q)$ and $\mathsf{purge}^{-1}(h) = \{q\}$. Then (1) follows trivially since both sides are equal to 1. To see (2), note that, by definition of σ',

$$\sigma'(h)(a) = \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h \cdot a) \mid \mathsf{purge}^{-1}(h)\right] = \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h \cdot a)\right]$$

since $h = f(q)$ here. Furthermore,

$$\begin{aligned}\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h \cdot a\right] &= \mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], a \mid h\right] \mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h\right] \\ &= \sigma'(h)(a)\end{aligned}$$

since $\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h\right] = 1$; which yields (2).

Last, assume that $\mathsf{stay} \in A'(f(q))$, that is $f(q)$ has the form s_D for some non-distinguishing MCEC D.

$$\begin{aligned}\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h \cdot \mathsf{stay}\right] &= \sigma'(h)(\mathsf{stay})\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h\right] \\ &= \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{purge}^{-1}(h)D^{\omega} \mid \mathsf{purge}^{-1}(h)] \\ &= \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{purge}^{-1}(h)D^{\omega}],\end{aligned}$$

which proves (3).

Assume now that $|h| > 1$, and let us write $h = h'ar$ for a history h', $a \in A'(\mathsf{last}(h'))$.

$$\begin{aligned}\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h'a\right] &= \mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h'a \mid h'\right] \mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h'\right] \\ &= \mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h'a \mid h'\right] \mathbb{P}^{\sigma}_{q}\left[M'[e], \mathsf{purge}^{-1}(h')\right] \\ &= \sigma'(h')(a)\mathbb{P}^{\sigma}_{q}\left[M'[e], \mathsf{purge}^{-1}(h')\right] \\ &= \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h'a) \mid \mathsf{purge}^{-1}(h')\right] \\ &\qquad \cdot \mathbb{P}^{\sigma}_{q}\left[M'[e], \mathsf{purge}^{-1}(h')\right], \\ &= \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h'a)\right],\end{aligned}$$

where we used the induction hypothesis to apply (1) on the second line. This proves (2).

Consider now $r \in Q'$.

$$\begin{aligned}\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h'ar\right] &= \mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h'ar \mid h'a\right]\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h'a\right] \\ &= \delta'_e(\mathsf{last}(h'), a)(r)\mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h'a)\right].\end{aligned}$$

We distinguish two cases. If $\mathsf{last}(h)$ does not have the form of s_D, then it also belongs to Q, $a \in A(\mathsf{last}(q))$, with $\delta_e(\mathsf{last}(h'), a)(r) = \delta'_e(\mathsf{last}(h'), a)(r)$. In this case, the above is equal to $\mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h'ar)\right]$. Assume now that $\mathsf{last}(h) = s_D$ for some non-distinguishing MCEC D, and that $a = F_{(r',a')}$ for some pair (r', a'). Then $\mathsf{purge}^{-1}(h'a)$ only contains histories that end at r', followed by action a'. We have, moreover, $\delta'_e(\mathsf{last}(h'), a)(r) = \sum_{q \in f^{-1}(r)} \delta_e(r', a')(q)$, by the definition of $\mathsf{purge}(M)$, so

$$\begin{aligned}\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h'ar\right] &= \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h'a)f^{-1}(r)\right] \\ &= \mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h'ar)\right].\end{aligned}$$

This proves (1).

Last, consider history h ending at some state s_D, and write

$$\begin{aligned}\mathbb{P}^{\sigma'}_{f(q)}\left[M'[e], h \cdot \mathsf{stay}\right] &= \mathbb{P}^{\sigma'}_{f(q)}\left[M[e], h \cdot \mathsf{stay} \mid h\right]\mathbb{P}^{\sigma'}_{f(q)}\left[M[e], h\right] \\ &= \mathbb{P}^{\sigma'}_{f(q)}\left[M[e], h \cdot \mathsf{stay} \mid h\right]\mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h)\right] \\ &= \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{purge}^{-1}(h)D^{\omega} \mid \mathsf{purge}^{-1}(h)]\mathbb{P}^{\sigma}_{q}\left[M[e], \mathsf{purge}^{-1}(h)\right] \\ &= \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{purge}^{-1}(h)D^{\omega}],\end{aligned}$$

where we used the induction hypothesis on the second line, and the definition of σ' on the third line. This proves (3).

We now show that $\mathbb{P}^{\sigma}_{q}(M, \varphi) \leq \mathbb{P}^{\sigma'}_{f(q)}(M', \varphi)$ follows from these properties. In fact, for all $e \in E$, one can write

$$\begin{aligned}\mathbb{P}^{\sigma'}_{f(q)}(M'[e], \varphi) = &\sum\nolimits_{\substack{D \in \mathsf{EC}(M[e]),\ \varphi\text{-winning} \\ \text{non-distinguishing MCEC}}} \mathbb{P}^{\sigma'}_{f(q)}(M'[e], (Q'A')^{*} \cdot s_D \cdot \mathsf{stay} \cdot q_{\mathsf{win}}) \\ &+ \sum\nolimits_{\substack{D \in \mathsf{EC}(M'[e]),\ \varphi\text{-winning} \\ D \neq \{(q_{\mathsf{win}}, \_)\}}} \mathbb{P}^{\sigma'}_{f(q)}(M'[e], \mathsf{Inf} = D),\end{aligned} \tag{4}$$

by separating winning end-components of M' into two: the winning absorbing state q_{win} reached via some s_D for a non-distinguishing MCEC of M, and any other end-component of M'.

For the first term of (4), we have, from the above properties of σ',

$$\begin{aligned}&\sum_{\substack{D \in \mathsf{EC}(M[e]),\ \varphi\text{-winning} \\ \text{non-distinguishing MCEC}}} \mathbb{P}^{\sigma'}_{f(q)}(M'[e], (Q'A')^{*} \cdot s_D \cdot \mathsf{stay} \cdot q_{\mathsf{win}}) \\ =& \sum_{\substack{D \in \mathsf{EC}(M[e]),\ \varphi\text{-winning} \\ \text{non-distinguishing MCEC}}} \sum_{\substack{D' \in \mathsf{EC}(M[e]) \\ D' \subseteq D}} \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{Inf} = D'] \\ \geq& \sum_{\substack{D \in \mathsf{EC}(M[e]),\ \varphi\text{-winning} \\ \text{non-distinguishing MCEC}}} \sum_{\substack{D' \in \mathsf{EC}(M[e]),\ \varphi\text{-winning} \\ D' \subseteq D}} \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{Inf} = D'] \\ =& \sum_{\substack{D \in \mathsf{EC}(M[e]),\ \varphi\text{-winning} \\ \text{non-distinguishing MCEC}}} \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{Inf} = D].\end{aligned}$$

For the second term of (4), let $D \in \mathsf{EC}(M') \setminus \{(q_{\mathsf{win}}, \_)\}$, and observe that $\mathsf{Inf} = D$ does not contain the action stay. Notice how we only have an inequality because σ might actually have a nonzero probability of realizing $\mathsf{Inf} = D'$ for some non-winning D' included in a winning D.

We established above that for all histories h of M' without the action stay, $\mathbb{P}^{\sigma'}_{f(q)}[M'[e], h] = \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{purge}^{-1}(h)]$, that is, cylinders generated by h and $\mathsf{purge}^{-1}(h)$ have the same probabilities in M' under σ', and, respectively, in M under σ. It follows that

$$\begin{aligned}
&\sum_{\substack{D \in \mathsf{EC}(M'[e]),\ \varphi\text{-winning} \\ D \neq \{(q_{\mathsf{win}}, \_)\}}} \mathbb{P}^{\sigma'}_{f(q)}[M'[e], \mathsf{Inf} = D] \\
=& \sum_{\substack{D \in \mathsf{EC}(M'[e]),\ \varphi\text{-winning} \\ D \neq \{(q_{\mathsf{win}}, \_)\}}} \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{purge}^{-1}(\mathsf{Inf} = D)] \\
=& \sum_{\substack{D \in \mathsf{EC}(M'[e]),\ \varphi\text{-winning} \\ D \neq \{(q_{\mathsf{win}}, \_)\}}} \sum_{D' \in \mathsf{EC}(M), f(D') = D} \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{Inf} = D'] \\
\geq& \sum_{\substack{D \in \mathsf{EC}(M'[e]),\ \varphi\text{-winning} \\ D \neq \{(q_{\mathsf{win}}, \_)\}}} \sum_{\substack{D' \in \mathsf{EC}(M), f(D') = D \\ D' \text{ is } \varphi\text{-winning}}} \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{Inf} = D'], \\
=& \sum_{\substack{D \in \mathsf{EC}(M[e]),\ \varphi\text{-winning} \\ \text{not a non-distinguishing MCEC}}} \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{Inf} = D].
\end{aligned}$$

where we extend the definition of f to state-action pairs so that $f(D')$ denotes an end-component of M'.

Combining these bounds on both terms of (4), we conclude

$$\begin{aligned}
\mathbb{P}^{\sigma'}_{f(q)}(M'[e], \varphi) \geq& \sum_{\substack{D \in \mathsf{EC}(M[e]),\ \varphi\text{-winning} \\ \text{non-distinguishing MCEC}}} \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{Inf} = D] \\
&+ \sum_{\substack{D \in \mathsf{EC}(M[e]),\ \varphi\text{-winning} \\ \text{not a non-distinguishing MCEC}}} \mathbb{P}^{\sigma}_{q}[M[e], \mathsf{Inf} = D] \\
&\geq \mathbb{P}^{\sigma}_{q}(M[e], \varphi).
\end{aligned}$$

◀

The following lemma is the dual, and shows that any strategy for $\mathsf{purge}(M)$ can be replicated in M, albeit with a bit more memory. The additional memory is required to implement behaviors inside non-distinguishing MCECs.

▶ **Lemma 20.** *Consider an MEMDP $M = \langle Q, A, (\delta_e)_{e \in E} \rangle$, and objective $\varphi = \mathsf{Parity}(p)$, and the map $f : Q \to Q'$ relating states of M and that of $\mathsf{purge}(M) = \langle Q', A', (\delta'_e)_{e \in E} \rangle$. For all states $q' \in Q'$ and strategies σ' for $\mathsf{purge}(M)$, and all $q \in f^{-1}(q')$, there exists a strategy σ with $\mathbb{P}^{\sigma}_{q}(M, \varphi) = \mathbb{P}^{\sigma'}_{q'}(\mathsf{purge}(M), \varphi)$. Furthermore, if σ' is a m-memory strategy, σ can be chosen to be a $(m + |Q||A|)$-memory strategy.*

Proof. Consider $q' \in Q'$, an m-memory strategy σ' for M', and $q \in f^{-1}(q')$, where m can be finite or infinite. We show that there exists an $(m + |Q||A|)$-memory strategy σ with $\mathbb{P}^{\sigma}_{q}(M, \varphi) = \mathbb{P}^{\sigma'}_{q'}(M', \varphi)$. We define σ as follows. Consider a history h of M.

- If $f(\mathsf{last}(h)) \in Q$, we let $\sigma(h) = \sigma'(\mathsf{purge}(h))$.

- Assume $f(\mathsf{last}(h)) = s_D$ for some non-distinguishing MCEC D. With probability $\sigma'(\mathsf{purge}(h))(\mathsf{stay})$, we let σ switch to a pure memoryless strategy that maximizes the probability of φ inside D (this strategy is independent from the environment). For each $F_{(q,a)}$, with probability $\sigma'(\mathsf{purge}(h))(F_{(q,a)})$ we let σ run a pure memoryless strategy until state q is reached (which happens probability 1), and from q take a.

The memory bound for σ is $m + |Q||A|$ where m is the memory size of σ', because inside each collapsed MCEC, and for each pair $F_{(q,a)}$, a pure memoryless strategy is executed until reaching q and taking action a.

By construction, for all histories h that start at q in M and end outside of non-distinguishing end-components, we have:

$$\mathbb{P}^{\sigma}_{q}(M[e], h) = \mathbb{P}^{\sigma'}_{f(q)}(M'[e], \mathsf{purge}(h)) \text{ for all environments } e \in E. \tag{5}$$

So if R denotes a measurable set of infinite runs of M such that for all $\rho \in R$, $\mathsf{purge}(\rho)$ is infinite (in other terms, ρ does not stay inside a non-distinguishing MCEC), then

$$\mathbb{P}^{\sigma}_{q}(M[e], R) = \mathbb{P}^{\sigma'}_{f(q)}(M'[e], \mathsf{purge}(R)), \tag{6}$$

writing $\mathsf{purge}(R) = \{\mathsf{purge}(\rho) \mid \rho \in R\}$.

Furthermore, for those histories $h = h'as$ where $q \in D$ is a non-distinguishing MCEC and $\mathsf{last}(h') \notin D$, we have:

$$\mathbb{P}^{\sigma}_{q}(M[e], h) = \mathbb{P}^{\sigma'}_{f(q)}(M'[e], \mathsf{purge}(h)) \text{ for all environments } e \in E.$$

Then, by definition of σ, for a non-distinguishing MCEC D,

$$\sum_{D' \in \mathsf{EC}(M), D' \subseteq D} \mathbb{P}^{\sigma}_{q}(M[e], \mathsf{Inf} = D') = \mathbb{P}^{\sigma'}_{f(q)}(M'[e], (Q'A')^{*} \cdot s_D \cdot \mathsf{stay}). \tag{7}$$

Observe that any end-component D of M that is not a non-distinguishing MCEC maps to an end-component of M'. We have for all $e \in E$,

$$\begin{aligned}
\mathbb{P}^{\sigma}_{q}(M[e], \varphi) = & \sum_{\substack{D \in \mathsf{EC}(M), \varphi\text{-winning} \\ \text{not a non-distinguishing MCEC}}} \mathbb{P}^{\sigma}_{q}(M[e], \mathsf{Inf} = D) \\
& + \sum_{\substack{D \in \mathsf{EC}(M), \varphi\text{-winning} \\ \text{non-distinguishing MCEC}}} \mathbb{P}^{\sigma}_{q}(M[e], \mathsf{Inf} = D) \\
= & \sum_{\substack{D \in \mathsf{EC}(M'), \varphi\text{-winning} \\ \text{not a non-distinguishing MCEC}}} \mathbb{P}^{\sigma'}_{f(q)}(M'[e], \mathsf{Inf} = D) \\
& + \sum_{\substack{D \in \mathsf{EC}(M'), \varphi\text{-winning} \\ \text{non-distinguishing MCEC}}} \mathbb{P}^{\sigma'}_{f(q)}(M'[e], (Q'A')^{*} \cdot s_D \cdot \mathsf{stay}) \\
= & \, \mathbb{P}^{\sigma'}_{f(q)}(M'[e], \varphi),
\end{aligned}$$

using (6) and (7). ◀

5.2 Learning While Playing

In this section, we show that after collapsing non-distinguishing MCECs, over n steps (for n large enough), with high probability, we either reach a MCEC (which is either distinguishing or trivial) or collect a large number of samples of distinguishing transitions whose empirical average is close to their mean. Intuitively, this means that either the knowledge can be improved after n steps using the collected samples while bounding the probability of error, or a MCEC is reached.

If the MCEC is distinguishing, the strategy can improve the knowledge as in Lemma 9, and if not, then the MCEC is trivial and there is a unique way to play. These results will be used in the next section to build a finite-memory strategy with approximately the same probability of winning, given any arbitrary strategy.

For a history h, let $|h|_{q,a}$ denote the number of occurrences of the state-action pair (q,a), and $|h|_{q,a,q'}$ the number of times these are followed by q', where $q' \in \mathsf{Supp}(\delta(q,a))$. For a distinguishing transition $t = (q,a,q')$, we say that a history h is a *bad (t,η)-classification* in MDP $M[e]$ if $\left|\frac{|h|_{q,a,q'}}{|h|_{q,a}} - \delta_e(q,a)(q')\right| \geq \eta/2$, that is the measured and theoretical frequency of t are too far apart. It is a *good (t,η)-classification* otherwise. Intuitively, over long histories, good classifications have high probability.

We first prove the following technical lemma, bounding the difference between the empirical average and the mean when sampling among a finite number of transitions, when the transitions to sample are chosen at each step by an *adversary*. This adversary corresponds to strategies in an MDP, is arbitrary, and can depend on the history and use randomization.

We state the following lemma for (single-environment) MDPs, and apply it to each environment in an MEMDP.

▶ **Lemma 21.** *Consider MDP M, state q_0, and $T = \{t_i = (q_i, a_i, q_i')\}_{1\leq i\leq k}$ a subset of transitions such that $(q_i,a_i) = (q_j,a_j)$ implies $q_i' = q_j'$ for all i,j. For all $\eta, \varepsilon > 0$, all $n_0 > \frac{k^3}{\varepsilon\eta^2}$, and any strategy σ with $\mathbb{P}^{\sigma}_{q_0}\left[\{h : \sum_{(q,a,q')\in T}|h_{q,a}| \geq n_0\}\right] = 1$, the set of histories h that satisfy the following conditions has probability at most ε:*

- $\sum_{(q,a,q')\in T}|h_{q,a}| \geq n_0$
- *there exists $1\leq i\leq k$ such that $|h|_{q_i,a_i} = \max_{i'}|h|_{q_{i'},a_{i'}}$ and h is a bad (t_i,η)-classification.*

Here the assumption on T simplifies the proofs since it means that for each state-action pair, we will be observing the frequency of a unique successor state. The lemma also requires that at least n_0 occurrences of T is visited with probability 1. This hypothesis ensures that we have enough samples to obtain a good approximation (that is, a good (t_i,η)-classification) with high probability (at least $1-\varepsilon$). In fact, if a strategy σ avoids visiting transitions from T, say, with probability 1/2, then it cannot ensure a good approximation with high probability because half the cases, there are just not enough samples of T.

The lemma is easy for $k = 1$. In fact, all trials are identical and independent, so one can use e.g. Hoeffding's inequality to derive a bound. When $k > 1$, trials are no longer independent since σ might react to the success or failure of a given transition to make its decisions in the future. In fact, the lemma is not trivial to prove due to the possible dependency between the trials.

Here is such a situation of dependency. Consider a state q from which action a leads to either to q_1 or q_2, each with probability 0.5, from which a deterministic transition comes back to q. Another action b from q deterministically loops back at q. Consider σ that picks (q,a) first. As long as we reach q_1, σ continues to pick (q,a). Whenever q_2 is reached, σ switches definitively to (q,b). Now the probability of observing (q,a,q_1) at step $n > 1$ depends on the result of the first $n-1$ trials. For example, conditioned on observing (q,a,q_1) on the first $n-1$ trials, the probability of observing (q,a,q_1) again is 0.5. But conditioned on not observing (q,a,q_1) on the $(n-1)$-th trial, this probability is 0. This shows that given such σ, the successive trials are not independent, and theorems such as Hoeffding's inequality cannot be applied.

In turns out that although the trials can be dependent, their covariance is 0. We exploit this observation to derive a good bound using Chebyshev's inequality:

▶ **Theorem 22** (Chebyshev's Inequality). *Let X be a random variable with mean μ, and standard deviation q. Then, for all $a > 0$, we have $\mathbb{P}[|X - \mu| \geq sa] \leq \frac{1}{a^2}$.*

This inequality clearly also applies if q is an upper bound on the standard deviation of X.

Proof of Lemma 21. We consider a slightly more abstract setting where there are k independent arms, each with a probability of success of p_i. In MDPs, each arm corresponds to a state-action pair (q_i, a_i) and it succeeds when reaching q'_i, with probability $p_i = \delta(q_i, a_i)(q'_i)$.

Consider a strategy σ that chooses, at each step, $i \in \{1, \ldots, k\}$, an arm to pull based on the full history and randomization. Consider $\varepsilon, \eta > 0$.

We model the problem as follows. For each $i \in \{1, \ldots, k\}$, define a sequence $X_1^{(i)}, X_2^{(i)}, \ldots$ of identical and independent Bernoulli variables with probability p_i. Let Choice_j denote the arm selected by σ at step j. At each step j, Choice_j selects an arm, and all types of arms are pulled. While Choice_j can depend on the history, $X_j^{(i)}$ does not depend on the history, and in particular on Choice_j.

Define the *weight* of arm i at step j as the following random variable.

$$\mathsf{wgt}_j^{(i)} = \begin{cases} X_j^{(i)} - p_i & \text{if } \mathsf{Choice}_j = i, \\ 0 & \text{otherwise.} \end{cases}$$

Define $\mathsf{wgt}_{\leq n}^{(i)} = \sum_{j=1}^{n} \mathsf{wgt}_j^{(i)}$, for any $n \geq 1$. Let us also define $\mathsf{occ}_j^{(i)} = 1$ iff $\mathsf{Choice}_j = i$, and $\mathsf{occ}_{\leq n}^{(i)} = \sum_{j=1}^{n} \mathsf{occ}_j^{(i)}$. Observe that

$$\mathsf{wgt}_{\leq n}^{(i)} = \sum_{1 \leq j \leq n, \mathsf{Choice}_j = i} X_j^{(i)} - \mathsf{occ}_{\leq n}^{(i)} p_i,$$

that is, this is the difference between the empirical sum and the mean of the sum of the subsequence of $X_j^{(i)}$ where $\mathsf{Choice}_j = i$.

Then $\frac{\mathsf{wgt}_{\leq n}^{(i)}}{\mathsf{occ}_{\leq n}^{(i)}}$ is the difference between the empirical average of the $X_j^{(i)}$ and p_i, assuming that $\mathsf{occ}_{\leq n}^{(i)} > 0$.

We have, by the definition of variance,

$$\mathbb{E}^\sigma[\mathsf{wgt}_j^{(i)}] = \mathbb{P}^\sigma[\mathsf{Choice}_j = i](p_i(1 - p_i) + (1 - p_i)(-p_i)) = 0,$$

so $\mathbb{E}^\sigma[\mathsf{wgt}_{\leq n}^{(i)}] = 0$ as well.

We are going to apply Theorem 22 on the variable $\mathsf{wgt}_{\leq n}^{(i)}$; so we need a bound on the variance of $\mathsf{wgt}_{\leq n}^{(i)}$. We show that $\mathbb{V}^\sigma[\mathsf{wgt}_{\leq n}^{(i)}] \leq n p_i (1 - p_i)$. We have

$$\mathbb{V}^\sigma[\mathsf{wgt}_{\leq n}^{(i)}] = \sum_{j=1}^{n} \mathbb{V}^\sigma[\mathsf{wgt}_j^{(i)}] + 2 \sum_{1 \leq j < j' \leq n} \mathsf{Cov}(\mathsf{wgt}_j^{(i)}, \mathsf{wgt}_{j'}^{(i)})$$

For each j, because $\mathbb{E}^\sigma[\mathsf{wgt}_j^{(i)}] = 0$, we have $\mathbb{V}^\sigma[\mathsf{wgt}_j^{(i)}] = \mathbb{E}^\sigma[(\mathsf{wgt}_j^{(i)})^2]$, which can be calculated as

$$\begin{aligned} &\mathbb{P}^\sigma[\mathsf{Choice}_j = i](p_i(1 - p_i)^2 + (1 - p_i)(-p_i)^2) \\ &= \mathbb{P}^\sigma[\mathsf{Choice}_j = i] p_i (1 - p_i)((1 - p_i) + p_i) \\ &\leq p_i(1 - p_i), \end{aligned}$$

so that the first term of the variance is at most $np_i(1-p_i)$.

Now, as noted above, $\mathsf{wgt}_j^{(i)}$ and $\mathsf{wgt}_{j'}^{(i)}$ are not independent variables since σ can choose the arm at step j' depending on the result of $\mathsf{wgt}_j^{(i)}$; we nevertheless show that the covariance is equal to 0. We have $\mathsf{Cov}(\mathsf{wgt}_j^{(i)}, \mathsf{wgt}_{j'}^{(i)}) = \mathbb{E}^\sigma[\mathsf{wgt}_j^{(i)} \cdot \mathsf{wgt}_{j'}^{(i)}] - \mathbb{E}^\sigma[\mathsf{wgt}_j^{(i)}][\mathsf{wgt}_{j'}^{(i)}]$ by definition of covariance; so this is equal to $\mathbb{E}^\sigma[\mathsf{wgt}_j^{(i)} \cdot \mathsf{wgt}_{j'}^{(i)}]$ which can be calculated as follows.

$$\begin{aligned}
&\mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1 \wedge X_{j'}^{(i)} = 1](1-p_i)^2\\
&+ \mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1 \wedge X_{j'}^{(i)} = 0](1-p_i)(-p_i)\\
&+ \mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 0 \wedge X_{j'}^{(i)} = 1](-p_i)(1-p_i)\\
&+ \mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 0 \wedge X_{j'}^{(i)} = 0](-p_i)^2.
\end{aligned}$$

Now $X_{j'}^{(i)}$ and the variables $\mathsf{Choice}_j, \mathsf{Choice}_{j'}, X_j^{(i)}$ are independent; in fact, the values of $\mathsf{Choice}_j, \mathsf{Choice}_{j'}$ cannot depend on $X_{j'}^{(i)}$ since the latter is revealed after $\mathsf{Choice}_j, \mathsf{Choice}_{j'}$. In contrast, $X_j^{(i)}$ and $\mathsf{Choice}_{j'}$ can be dependent since the latter can depend on the value of $X_j^{(i)}$.

We can rewrite $\mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1 \wedge X_{j'}^{(i)} = 1]$ as follows.

$$\begin{aligned}
&\mathbb{P}^\sigma[X_{j'}^{(i)} = 1 \mid \mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1]\mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1]\\
&= \mathbb{P}^\sigma[X_{j'}^{(i)} = 1]\mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1]\\
&= p_i\mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1]
\end{aligned}$$

by independence.

Applying this to all four terms, $\mathbb{E}^\sigma[\mathsf{wgt}_j^{(i)} \cdot \mathsf{wgt}_{j'}^{(i)}]$ can be written as

$$\begin{aligned}
&\mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1]p_i(1-p_i)^2\\
&\quad+ \mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1](1-p_i)(1-p_i)(-p_i)\\
&\quad+ \mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 0]p_i(-p_i)(1-p_i)\\
&\quad+ \mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 0](1-p_i)(-p_i)^2,\\
&=\mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 1](p_i(1-p_i)^2 + (1-p_i)(1-p_i)(-p_i))\\
&\quad+ \mathbb{P}^\sigma[\mathsf{Choice}_j = i \wedge \mathsf{Choice}_{j'} = i \wedge X_j^{(i)} = 0](p_i(-p_i)(1-p_i) + (1-p_i)(-p_i)^2)\\
&=0.
\end{aligned}$$

So all covariance terms are 0, and we have $\mathbb{V}^\sigma[\mathsf{wgt}_{\le n}^{(i)}] \le np_i(1-p_i)$.

We now apply Theorem 22: For all $1 \le i \le k$, and for all $a > 0$,

$$\mathbb{P}^\sigma\left[\left|\mathsf{wgt}_{\le n}^{(i)}\right| \ge a\sqrt{np_i(1-p_i)}\right] \le \frac{1}{a^2}.$$

Using $\mathbb{P}[X \cup Y] = \mathbb{P}[X] + \mathbb{P}[Y] - \mathbb{P}[X \cdot Y]$, it follows that

$$\mathbb{P}^\sigma\left[\exists i, \left|\mathsf{wgt}_{\le n}^{(i)}\right| \ge a\sqrt{np_i(1-p_i)}\right] \le \frac{k}{a^2}.$$

We have,

$$\mathbb{P}^\sigma\left[\exists i, \mathsf{occ}_{\le n}^{(i)} = \max_{i'} \mathsf{occ}_{\le n}^{(i')} \wedge \left|\frac{\mathsf{wgt}_{\le n}^{(i)}}{\mathsf{occ}_{\le n}^{(i)}}\right| \ge \frac{a\sqrt{np_i(1-p_i)}}{\mathsf{occ}_{\le n}^{(i)}}\right] \le \frac{k}{a^2}.$$

Here we divided the inequality by $\mathsf{occ}^{(i)}_{\leq n}$ (since for $n > 0$, $\max_{i'} \mathsf{occ}^{(i')}_{\leq n} > 0$); moreover, the probability bound holds since each event has become smaller.

Notice that for all $n > 0$, $\sum_{i=1}^{k} \mathsf{occ}^{(i)}_{\leq n} = n$ since σ picks one of the arms at each step; so $\max_{i'} \mathsf{occ}^{(i')}_{\leq n} \geq n/k$ with probability 1. We get

$$\mathbb{P}^{\sigma}\left[\exists i, \mathsf{occ}^{(i)}_{\leq n} = \max_{i'} \mathsf{occ}^{(i')}_{\leq n} \wedge \left|\frac{\mathsf{wgt}^{(i)}_{\leq n}}{\mathsf{occ}^{(i)}_{\leq n}}\right| \geq \frac{ak\sqrt{p_i(1-p_i)}}{\sqrt{n}}\right] \leq \frac{k}{a^2}.$$

Now, given $\varepsilon, \eta > 0$, we pick $a = \sqrt{k/\varepsilon}$ so that $k/a^2 \leq \varepsilon$; and then n large enough so that $\frac{ak\sqrt{p_i(1-p_i)}}{\sqrt{n}} \leq \eta/2$; this means it suffices to pick n such that $\max_{1\leq i\leq k}\left(\frac{ak\sqrt{p_i(1-p_i)}}{\eta/2}\right)^2 \leq n$, so $(\frac{ak}{\eta/2})^2 = \frac{4k^3}{\varepsilon\eta^2} \leq n$ suffices. ◀

We use Lemma 21 to prove that in MEMDPs without non-trivial and non-distinguishing MCECs (for example, obtained by $\mathsf{purge}(\cdot)$), after n steps, we either reach a MCEC or collect a large number of samples of distinguishing transitions whose empirical average is close to their mean. Given MEMDP M, let T_M denote a set obtained by selecting one distinguishing transition (q, a, q') for each state-action pair (q, a) whose probability distribution differs in a pair of different environments. We select one representative distinguishing transition (q, a, q') for each pair (q, a) because this simplifies the calculations. We let $|h|_{T_M} = \sum_{(q,a,q')\in T_M} |h|_{q,a}$.

Let us fix η as follows

$$\eta < \frac{1}{2}\min\left(\{|\delta_e(q,a)(q') - \delta_f(q,a)(q')| \mid e, f \in E, q, q' \in Q, a \in A\} \setminus \{0\}\right).$$

Let us define the set of *good histories with* n_0 *samples*, denoted Good_{n_0}, as the set of histories h satisfying

- $|h|_{T_M} \geq n_0$,
- for all $t = (q, a, \_) \in T_M$ satisfying $|h|_{q,a} = \max_{(q',a',\cdot)\in T_M} |h|_{q',a'}$, h is a good (t, η)-classification.

▶ **Lemma 23.** *Consider an MEMDP M whose only non-distinguishing MCECs are trivial, and fix $\varepsilon > 0$, Let $n_0 = \lceil \frac{2(|Q||A|)^3}{\varepsilon\eta^2} \rceil$, and $n \geq 2p^{-2|Q|}\max(\log(\frac{4}{\varepsilon}), n_0)$ where p is the smallest nonzero probability that appears in M. Then, from any starting state, and under any strategy, with probability at least $1-\varepsilon$, within n steps, the history either visits a MCEC (distinguishing or trivial), or belongs to Good_{n_0}.*

Proof. We show that in all $M[e]$, under any strategy σ, from every state q_0, there is a path of size at most $|Q|$ compatible with the strategy that reaches a MCEC or a distinguishing transition. Consider first the case of a pure strategy σ. To prove this, towards a contraction, assume that MCECs and distinguishing transitions are not visited within $|Q|$ steps under σ. Consider the execution tree that starts at q_0 in M under σ: this is a tree labeled by Q, in which the children of a given node at history h are labeled by all possible successors $\mathsf{Supp}(\delta_e(\mathsf{last}(h), \sigma(h)))$ for some $e \in E$. Since all transitions are non-distinguishing, the choice of e is irrelevant here. We build this tree and cut each branch whenever a MCEC or a distinguishing transition is seen, or a state is repeated. Since we assumed that MCECs and distinguishing transitions are not reachable under σ, all branches of this tree are cut only when a state is repeated. It follows that the set of states in this tree, together with the actions prescribed by σ from these histories form a closed set of states. But then a strongly-connected subset must exist, which is a non-distinguishing CEC. This is thus included in a MCEC, contradicting our assumption.

If σ is pure, then in all $M[e]$, from every history, there is a probability of at least $p^{|Q|}$ of either taking a distinguishing transition, or visiting a MCEC within $|Q|$ steps, independently of the current state. If σ is randomized, then the probability of such a single run can be smaller since σ might assign small probabilities to its actions. In this case, since we are only interested in the behaviors in the first n steps, we can see σ as a *mixed* strategy which consists in randomly choosing among a set of pure strategies that stop after n steps. Since the above argument can be applied to each pure strategy in the support of σ (when σ is seen as a mixed strategy), it follows that under σ, there is a probability of at least $p^{|Q|}$ of taking a distinguishing transition or visiting a MCEC within the next $|Q|$ steps, as well.

Viewing runs as the concatenation of finite segments of size $|Q|$, we call each such segment a trial. Consider the random Bernoulli variables $X_1, X_2, \ldots$ such that the value of X_i is 1 iff a MCEC or a distinguishing transition is visited at the i-th trial.

So by Hoeffding's inequality, for all states q_0, $n > 0$ and $t > 0$, [2]

$$\mathbb{P}^{\sigma}_{q_0}\left[\sum_{i=1}^{n} X_i \leq \sum_{i=1}^{n} \mathbb{E}^{\sigma}[X_i] - t\right] \leq 2e^{-2\frac{t^2}{n}}.$$

Given $n > 0$, we choose here $t = np^{|Q|}/2$. This yields,

$$\mathbb{P}^{\sigma}_{q_0}\left[\sum_{i=1}^{n} X_i \leq \sum_{i=1}^{n} \mathbb{E}^{\sigma}[X_i] - np^{|Q|}/2\right] \leq 2e^{-2\frac{n^2p^{2|Q|}}{4n}} \leq \varepsilon/2,$$

which is the case since, by taking the log of both sides,

$$\begin{aligned} -\frac{np^{2|Q|}}{2} &\leq \log(\varepsilon/4) \\ &\Leftrightarrow n \geq 2\log(4/\varepsilon)p^{-2|Q|}. \end{aligned}$$

Because $\mathbb{E}^{\sigma}(X_i) \geq p^{|Q|}$, $\sum_{i=1}^{n} \mathbb{E}^{\sigma}[X_i] \geq np^{|Q|}$. This means that with probability at least $1-\varepsilon/2$, $\sum_{i=1}^{n} X_i \geq np^{|Q|}/2$. As $n \geq 2\lceil\frac{2(|Q||A|)^3}{\varepsilon\eta^2}\rceil p^{-2|Q|}$, we have $\sum_{i=1}^{n} X_i \geq \lceil\frac{2(|Q||A|)^3}{\varepsilon\eta^2}\rceil$, that is, with probability at least $1-\varepsilon/2$, either a MCEC or $\lceil\frac{2(|Q||A|)^3}{\varepsilon\eta^2}\rceil$ occurrences of distinguishing transitions are seen (which can be good or bad classifications).

Let us write $n_0 = \lceil\frac{2(|Q||A|)^3}{\varepsilon\eta^2}\rceil$. It remains to bound the probability of visiting either a MCEC or Good_{n_0}. Let us define a tree-shaped MDP M_n from M as follows. First, we unfold M by stopping each branch either when a MCEC is reached, or after n steps. Then, each leaf that belongs to a MCEC is extended with fresh states and transitions so that the branch contains n_0 instances of distinguishing transitions. More precisely, we pick some distinguishing transition (q, a, q') of M, and extend a given leaf l_0 of M_n as follows. The only enabled action at l_0 is a, and it goes to l_0' with probability $\delta_e(q, a, q')$ to l_0' in $M_n[e]$, and to

[2] Note that Hoeffding's inequality requires an independent sequence of random variables which is not the case of the X_i's. We can nevertheless still apply this inequality here using a coupling argument: Define U_i as a sequence of independent and continuous variables uniformly distributed over $[0, 1]$. Define the Bernoulli variable $Y_i = 1$ iff $U_i \leq p^{|Q|}$. Furthermore, define the sequence of Bernoulli variables $\tilde{X}_1, \tilde{X}_2, \ldots$ inductively, by $\tilde{X}_i = 1$ iff $U_i \leq \mathbb{P}(X_i = 1 \mid X_1 = \tilde{X}_1, \ldots, X_{i-1} = \tilde{X}_{i-1})$. Because $\mathbb{P}(X_i = 1 \mid h) \geq p^{|Q|}$ regardless of the history h, we have $Y_i \leq \tilde{X}_i$. Furthermore, $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(\tilde{X}_1 = x_1, \ldots, \tilde{X}_n = x_n)$ for all $x_1, \ldots, x_n \in \{0, 1\}$. It follows that Hoeffding's inequality can be applied on the i.i.d. sequence Y_i, and we get for all $A > 0$, $\mathbb{P}[\sum_i X_i \leq A] \leq \mathbb{P}[\sum_i Y_i \leq A]$.

l''_0 with probability $1 - \delta_e(q, a, q')$; and both l'_0, l''_0 deterministically go to l_1. We repeat this until n_0 occurrences of distinguishing transitions are obtained. Last, all leafs are made into absorbing states.

Let $\diamond$CEC denote the set of histories that reach a MCEC. For all $e \in E$,

$$\mathbb{P}^{\sigma}_{q_0}[M[e], \diamond\mathrm{CEC} \vee \mathsf{Good}_{n_0}] \geq \mathbb{P}^{\sigma}_{q'_0}[M_n[e], \mathsf{Good}_{n_0}]$$

where q'_0 is the root of $M_n[e]$, since MCECs are replaced with a gadget that might not satisfy Good_{n_0} with probability 1.

As an additional step, we obtain M'_n by modifying M_n as follows: we extend each leaf whose branch does not contain n_0 occurrences of distinguishing transitions (nor visit a MCEC), by adding fresh states and transitions as described above so that a total of n_0 distinguishing transitions is obtained at each branch. We get for all $e \in E$.

$$\mathbb{P}^{\sigma}_{q'_0}[M_n[e], \mathsf{Good}_{n_0}] \geq \mathbb{P}^{\sigma}_{q'_0}[M'_n[e], \mathsf{Good}_{n_0}] - \varepsilon/2$$

since the probability of the modified branches was shown to be at most ε above.

Now, by construction, for all strategies σ and $e \in E$, n_0 occurrences of distinguishing transitions are seen in $M'_n[e]$ with probability 1. By Lemma 21 with $k = |Q||A|$, applied for $\varepsilon/2$, we get

$$\mathbb{P}^{\sigma}_{q'_0}[M'_n[e], \mathsf{Good}_{n_0}] \geq 1 - \varepsilon/2.$$

It follows that $\mathbb{P}^{\sigma}_{q}[M[e], \diamond\mathrm{CEC} \vee \mathsf{Good}_{n_0}] \geq \mathbb{P}^{\sigma}_{q'_0}[M'_n[e], \mathsf{Good}_{n_0}] \geq 1 - \varepsilon$ for all $e \in E$, as required. ◀

5.3 Constructing Approximate Finite-Memory Strategies

We are now ready to construct a finite-memory strategy that approximates an arbitrary strategy σ. We construct a finite-memory strategy for $\mathsf{purge}(M)$ and then transfer it to M using Lemmas 19-20. The finite-memory strategy we construct consists in imitating the strategy σ for n steps, where n is defined in Lemma 23. Because all nontrivial MCECs of $\mathsf{purge}(M)$ are distinguishing, when we play for n steps, with high probability, we either visit a trivial MCEC (which is either winning for all environments or losing for all environments), or reach a distinguishing MCEC, or observe enough samples of distinguishing transitions. The strategy is extended arbitrarily in trivial MCECs. Inside distinguishing MCECs, it gathers samples of distinguishing transitions as in Lemma 9, which improves the knowledge (with an arbitrarily small probability of error). The knowledge is also correctly improved with high probability if enough samples are gathered outside of MCECs. In both cases, the strategy switches to a finite-memory strategy for the improved knowledge constructed recursively for smaller sets of environments.

Lemma 24 formalizes this reasoning and gives a bound N on the memory of the resulting strategy. In the memory bound, the term $\lceil 8\frac{\log(8/\varepsilon)}{\eta^2}^2\rceil$ comes from the application of Lemma 9 for distinguishing MCECs for each subset of E; and the term $(2|Q|)^{n(|E|+1)}$ corresponds to the recursive analysis, since the strategy is defined inductively for each subset of E.

▶ **Lemma 24.** *Consider an MEMDP $M = \langle Q, A, (\delta_e)_{e\in E}\rangle$, state $q \in Q$, parity objective φ. For all strategies σ, and $\varepsilon > 0$, there exists a strategy σ' using at most $N = (2|Q|)^{n(|E|+1)}|A|\lceil 8\frac{\log(8/\varepsilon)}{\eta^2}^2\rceil$ memory where $n = \left\lceil 2p^{-2|Q|}\max(\frac{8(|Q||A|)^3}{\varepsilon\eta^2}, \log(16/\varepsilon))\right\rceil$, with p the smallest nonzero probability in M, and that satisfies $\mathbb{P}^{\sigma'}_{q}(M, \varphi) \geq \mathbb{P}^{\sigma}_{q}(M, \varphi) - \varepsilon$.*

Proof. Given $\varepsilon > 0$, let

$$n_0 = \left\lceil \frac{8(|Q||A|)^3}{\varepsilon\eta^2} \right\rceil,$$
$$n = \left\lceil 2p^{-2|Q|} \max(n_0, \log(16/\varepsilon)) \right\rceil,$$

Notice that the bounds on n and n_0 come from Lemma 23 applied for $\varepsilon/4$. Define the sequence $(g_i)_{i\geq 1}$ by $g_1 = 1$, and $g_i = \alpha(g_{i-1} + \beta) + \gamma$ where $\alpha = 2|Q|^n$, $\beta = \lceil 8\frac{\log(8/\varepsilon)}{\eta^2}^2\rceil$, and $\gamma = |Q||A|$. Note that we have, for $i > 1$, $g_i = \alpha^{i-1} + (\gamma + \alpha\beta)(\frac{\alpha^{i-1}-1}{\alpha-1})$. Observe that $g_i \leq \alpha^{i-1}(1 + \gamma + \alpha\beta)$.

We prove, by induction on $|E|$, that for all states q, strategies σ, there exists a $g_{|E|}$-memory strategy σ' such that $\mathbb{P}_q^{\sigma'}(M, \varphi) \geq \mathbb{P}_q^{\sigma}(M, \varphi) - \varepsilon$.

We have $g_{|E|} \leq \alpha^{|E|-1}(1 + \gamma + \alpha\beta) \leq \alpha^{|E|}(1 + |Q||A| + 2|Q|^n\beta) \leq \alpha^{|E|}(3|Q|^n|A|\beta) \leq \alpha^{|E|}(2\alpha|A|\beta)$, which is at most $(2|Q|)^{n(|E|+1)}|A|\lceil 8\frac{\log(8/\varepsilon)}{\eta^2}^2\rceil$, and proves the lemma.

The base case $|E| = 1$ is obvious since there exists optimal memoryless strategies for parity objectives in MDPs. Assume $|E| \geq 2$.

Let $M' = \mathsf{purge}(M)$ and σ' be given by Lemma 19 such that $\mathbb{P}_q^{\sigma}(M, \varphi) \leq \mathbb{P}_{q'}^{\sigma'}(M', \varphi)$ where $q' = f(q)$. We prove the property for M' and transfer the result back to M using Lemma 20. More precisely, we show below that there exists a $(g_{|E|} - \gamma)$-memory strategy σ'' with $\mathbb{P}_{q'}^{\sigma''}(M', \varphi) \geq \mathbb{P}_{q'}^{\sigma'}(M', \varphi) - \varepsilon$. It follows, by Lemma 19, that there exists a $g_{|E|}$-memory strategy σ''' for M such that

$$\mathbb{P}_q^{\sigma'''}(M, \varphi) = \mathbb{P}_{q'}^{\sigma''}(M', \varphi) \geq \mathbb{P}_{q'}^{\sigma'}(M', \varphi) - \varepsilon \geq \mathbb{P}_q^{\sigma}(M, \varphi) - \varepsilon$$

which proves the result.

We construct σ'' by imitating σ' for n steps, and stopping if a MCEC is reached (thus, either trivial or distinguishing, by Lemma 17). More precisely, consider history h in M' that starts at q'. We define $\sigma''(h) = \sigma'(h)$, except in the following cases where σ'' switches to a strategy as described below:

1. If $\mathsf{last}(h)$ belongs to a trivial MCEC, then σ'' is memoryless from that history (as there is only one possible action to choose). Notice that $\mathbb{P}_{q'}^{\sigma''}(M', \varphi \mid h) = \mathbb{P}_{q'}^{\sigma'}(M', \varphi \mid h)$ since this MCEC is either winning or losing with probability 1, in each $e \in E$.
2. Assume $\mathsf{last}(h)$ belongs to a distinguishing MCEC D with partition (K_1, K_2). Let $\vec{\beta} = \mathbb{P}^{\sigma'}(M', \varphi \mid h)$, the probability values achieved from history h under strategy σ' starting with history h. One can define a strategy σ'_h such that $\vec{\beta} = \mathbb{P}_{\mathsf{last}(h)}^{\sigma'_h}(M', \varphi)$, by $\sigma'_h : h' \mapsto \sigma'(h \cdot h')$. By induction applied to M', $\mathsf{last}(h)$, σ'_h, environment set K_i, and $\varepsilon/8$, there exist $g_{|K_i|}$-memory strategies σ_i, with $\mathbb{P}_{\mathsf{last}(h)}^{\sigma_i}(M'[K_i], \varphi) \geq \vec{\beta}|_{K_i} - \varepsilon/8$. We apply Lemma 9 to build strategy σ''_h satisfying the following:
$$\mathbb{P}_{\mathsf{last}(h)}^{\sigma''_h}(M'[e], \varphi) \geq \mathbb{P}_{\mathsf{last}(h)}^{\sigma_i}(M'[e], \varphi) - \varepsilon/4 \text{ for all environments } e \in K_i. \quad (8)$$
At h, we let σ'' switch to σ''_h. It follows that $\mathbb{P}_{q'}^{\sigma''}(M', \varphi \mid h) \geq \mathbb{P}_{q'}^{\sigma'}(M', \varphi \mid h) - \varepsilon/4$.
3. Assume that h contains n_0 occurrences of distinguishing state-action pairs, that is, $|h|_{T_{M'}} = n_0$. Let $(q, a, q') \in T_M$ be a distinguishing transition with the largest number of occurrences in h; and let (K_1, K_2) be the partition of E induced by this transition. For each $i = 1, 2$, let σ_i be the $g_{|K_i|}$-memory strategy given by induction hypothesis applied to M', state $\mathsf{last}(h)$, environment set K_i, bound $\varepsilon/4$, and strategy $\sigma'_h : h' \mapsto \sigma'(h \cdot h')$ that achieves $\mathbb{P}_{\mathsf{last}(h)}^{\sigma_i}(M'[K_i], \varphi) \geq \mathbb{P}_{\mathsf{last}(h)}^{\sigma'_h}(M'[K_i], \varphi) - \varepsilon/4$ for each $i = 1, 2$. We let σ'' switch to:

- σ_1 if $\left|\frac{|h|_{q,a,q'}}{|h|_{q,a}} - \delta_e(q,a)(q')\right| < \eta/2$ for some $e \in K_1$,
- σ_2 otherwise.

The above shows that if h is a good classification in e, then $\mathbb{P}_{q'}^{\sigma''}(M'[e], \varphi \mid h) \geq \mathbb{P}_{q'}^{\sigma'}(M'[e], \varphi \mid h) - \varepsilon/4$.

4. If $|h| = n$ and none of the above applies, then σ'' switches to an arbitrary memoryless strategy. These histories that satisfy case 4 has probability at most $\varepsilon/4$ by Lemma 23.

Let us show that $\mathbb{P}_{q'}^{\sigma''}(M', \varphi) \geq \mathbb{P}_{q'}^{\sigma'}(M', \varphi) - \varepsilon$. To prove this, we distinguish histories h according to the cases above, and relate $\mathbb{P}_{q'}^{\sigma''}(M', \varphi \mid h)$ and $\mathbb{P}_{q'}^{\sigma'}(M', \varphi \mid h)$, and bound the probability of some histories h.

Let us write

$$\begin{aligned}\mathbb{P}_{q'}^{\sigma''}(M'[e], \varphi) = &\sum_{h\text{: case 1}} \mathbb{P}_{q'}^{\sigma''}(M'[e], \varphi, h) + \sum_{h \text{ case 2}} \mathbb{P}_{q'}^{\sigma''}(M'[e], h)\mathbb{P}_{q'}^{\sigma''}(M'[e], \varphi \mid h)\\ &+ \sum_{\substack{h\text{: case 3}\\ \text{bad classification}}} \mathbb{P}_{q'}^{\sigma''}(M'[e], h)\mathbb{P}_{q'}^{\sigma''}(M'[e], \varphi \mid h)\\ &+ \sum_{\substack{h\text{: case 3}\\ \text{good classification}}} \mathbb{P}_{q'}^{\sigma''}(M'[e], h)\mathbb{P}_{q'}^{\sigma''}(M'[e], \varphi \mid h)\\ &+ \sum_{h\text{: case 4}} \mathbb{P}_{q'}^{\sigma''}(M'[e], h)\mathbb{P}_{q'}^{\sigma''}(M'[e], \varphi \mid h).\end{aligned}$$

Since $\mathbb{P}_{q'}^{\sigma''}(M', h) = \mathbb{P}_{q'}^{\sigma'}(M', h)$ for histories satisfying any of the cases (because σ'' imitates σ' until such a case occurs), and because the terms $\mathbb{P}_{q'}^{\sigma''}(M', h)$ at the second and forth lines are each at most $\varepsilon/4$, using the cases above, we get

$$\begin{aligned}\mathbb{P}_{q'}^{\sigma''}(M'[e], \varphi) \geq &\sum_{h\text{:case 1}} \mathbb{P}_{q'}^{\sigma'}(M'[e], \varphi, h) + \sum_{h\text{: case 2}} \mathbb{P}_{q'}^{\sigma'}(M'[e], h)(\mathbb{P}_{q'}^{\sigma'}(M'[e], \varphi | h) - \varepsilon/4)\\ &+ \left(\sum_{\substack{h\text{: case 3}\\ \text{bad classification}}} \mathbb{P}_{q'}^{\sigma'}(M'[e], h)\mathbb{P}_{q'}^{\sigma'}(M'[e], \varphi \mid h) \right) - \varepsilon/4\\ &+ \sum_{\substack{h\text{: case 3}\\ \text{good classification}}} \mathbb{P}_{q'}^{\sigma'}(M'[e], h)(\mathbb{P}_{q'}^{\sigma'}(M'[e], \varphi \mid h) - \varepsilon/4)\\ &+ \left(\sum_{h\text{: case 4}} \mathbb{P}_{q'}^{\sigma''}(M'[e], h)\mathbb{P}_{q'}^{\sigma''}(M'[e], \varphi \mid h) \right) - \varepsilon/4.\\ &\geq \mathbb{P}_{q'}^{\sigma'}(M'[e], \varphi) - \varepsilon.\end{aligned}$$

Last, we argue that σ'' uses memory of size $g_{|E|}$. Strategy σ'' must store the histories until one of the four cases occur: this happens in at most n steps, which means $|Q|^n$ memory is required for this phase. In addition, for each history of case 2, $g_{|E|-1} + g_{|E|-1} + 8\frac{\log(8/\varepsilon)}{\eta^2}^2$ memory states are needed by Lemma 9; where the terms $g_{|E|-1}$ are upper bounds on the memory requirement of the strategies to which we switch, given by induction. Case 3 does not require additional memory since the decision is made depending on the current history, which is already in the memory. In total, we thus need $|Q|^n(2g_{|E|-1} + \beta)$ memory states, which is at most $\alpha(g_{|E|-1} + \beta) = g_{|E|} - \gamma$. ◀

5.4 Approximation Algorithm

We now provide a procedure solving the gap problem with threshold α for parity objectives in MEMDPs. Informally, given bound N, the procedure *guesses* an N-memory strategy by solving a set of polynomial constraints over the reals, and checks that the strategy ensures winning with probability at least $\alpha - \varepsilon$ in all environments. We first give the construction for reachability, then explain the extension to parity conditions.

Reachability in MDPs Let us start by recalling the linear constraints that characterize reachability probabilities in single-environment MDPs under memoryless strategies. Consider an MDP $M = \langle Q, A, \delta \rangle$ and objective $\mathsf{Reach}(T)$. Let $Q^{\mathsf{no}} \subseteq Q$, and $Q^{?} = Q \setminus (Q^{\mathsf{no}} \cup T)$. Q^{no} will be the set of states from which the reachability probability is 0; it is necessary to make sure all such states are in Q^{no} so that the equation given below has a unique solution. Define the unknown x_q representing the probability of reaching T from q under the strategy that is being guessed, and $p_q(a)$ the probability of the strategy to pick action a from q, for $a \in A_q$. Consider the following constraints:

$$\begin{aligned}
&x_q = 0 && \text{for all } q \in Q^{\mathsf{no}},\\
&x_q = 1 && \text{for all } q \in T,\\
&x_q = \textstyle\sum_{a \in A_q} p_q(a) \cdot \sum_{q' \in Q} \delta(q,a,q') \cdot x_{q'} && \text{for all } q \in Q^{?},\\
&0 \leq x_q \leq 1 \text{ and } 0 \leq p_q(a) \leq 1 && \text{for all } q \in Q, a \in A_q,\\
&\textstyle\sum_{a \in A_q} p_q(a) = 1 && \text{for all } q \in Q.
\end{aligned} \tag{9}$$

Any solution $(\vec{x}, \vec{p})$ of (9) yields a strategy $\sigma^{\vec{p}}$, which is defined as picking action a from state q with probability $p_q(a)$. The following theorem shows that $\vec{x}$ does capture the reachability probabilities of $\sigma^{\vec{p}}$, provided that Q^{no} is the set of states from which the reachability probability is 0.

▶ **Theorem 25** (Theorem 10.19, [3]). *Consider any subset $Q^{no} \subseteq Q$, and a solution $(\vec{x}, \vec{p})$ of (9). If for all states $q \in Q^{no}$, $\mathbb{P}_q^{\sigma^{\vec{p}}}[M, \mathsf{Reach}(T)] = 0$, then for all $q \in Q$, $x_q = \mathbb{P}_q^{\sigma^{\vec{p}}}[M, \mathsf{Reach}(T)]$. Conversely, for any memoryless strategy τ, if Q^{no} denotes the set of states q with $\mathbb{P}_q^{\tau}[M, \mathsf{Reach}(T)] = 0$, then (9) has a unique solution $(\vec{x}, \vec{p})$ where $\tau = \sigma^{\vec{p}}$, and $x_q = \mathbb{P}_q^{\tau}[M, \mathsf{Reach}(T)]$ for all $q \in Q$.*

Finite-Memory Reachability in MEMDPs We now show how to solve the gap problem for an instance of the quantitative reachability problem for MEMDPs. Consider MEMDP $M = \langle Q, A, (\delta_e)_{e \in E} \rangle$, objective $\mathsf{Reach}(T)$, a memory bound N, an initial state q_0, and bounds $\varepsilon, \alpha > 0$. We want to check whether there exists a strategy σ such that, for all environments $e \in E$, we have $\mathbb{P}_{q_0}^{\sigma}[M[e], \mathsf{Reach}(T)] \geq \alpha$, or whether for all σ, there exists an environment $e \in E$ with $\mathbb{P}_{q_0}^{\sigma}[M[e], \mathsf{Reach}(T)] < \alpha - \varepsilon$.

We guess a memoryless randomized strategy on combined states (q,i) for $q \in Q$ and $0 \leq i < N$, which correspond to N-memory strategies on M. In the sequel, we write $[N] = \{0, 1, \ldots, N-1\}$. We define the unknown variable $x_{q,i}^e$ for each $e \in E$, and combined state (q,i) representing the probability of reaching T from state q and memory value i in $M[e]$, under the strategy that is being guessed. Furthermore, define $p_{q,i}(a, i')$ for each action $a \in A_q$ and $i' \in [N]$, as the unknown representing the probability of the strategy picking action a from (q,i) and updating the memory value to i'.

Consider subsets $Q_e^{\mathsf{no}} \subseteq Q$ for each $e \in E$, and let $Q_e^{?} = Q \setminus (Q_e^{\mathsf{no}} \cup T_e)$. We write the following constraints in a slightly more general setting, where a possibly different target set T_e is considered for each environment e (this will be useful when generalizing to parity conditions below):

$$\begin{array}{ll}
x^e_{q,i} = 0 & \text{for all } e \in E, q \in Q^{\mathsf{no}}_e, i \in [N], \\
x^e_{q,i} = 1 & \text{for all } e \in E, q \in T_e, i \in [N], \\
x^e_{q,i} = \sum_{a \in A_q} p_{q,i}(a,i') \cdot \sum_{q' \in Q} \delta_e(q,a,q') \cdot x^e_{q',i'} & \text{for all } e \in E, q \in Q^?_e, i \in [N], \\
0 \leq x^e_{q,i} \leq 1 & \text{for all } e \in E, q \in Q, i \in [N], \\
0 \leq p_{q,i}(a,i') \leq 1 & \text{for all } q \in Q, a \in A_q, i, i' \in [N], \\
\sum_{a \in A_q} \sum_{i' \in [N]} p_{q,i}(a,i') = 1 & \text{for all } q \in Q, i \in [N], \\
x^e_{q_0} \geq \alpha - \varepsilon & \text{for all } e \in E.
\end{array} \tag{10}$$

Notice how the choice of the action and memory updates $p_{q,i}$ does not depend on the environment. The constraints (10) simply combine $|E|$ copies of (9) over a state space augmented with N memory values. In addition we added the constraints $x^e_{q_0} \geq \alpha - \varepsilon$ for all $e \in E$, which restrict the solution sets to those strategies that ensure the threshold $\alpha - \varepsilon$.

The Gap Problem for Reachability The full procedure is as follows. We let $T_e = T$ for all $e \in E$. Let N be the memory bound given in Lemma 24.

We enumerate all possibles choices for the sets Q^{no}_e. For each choice $(Q^{\mathsf{no}}_e)_{e \in E}$, we solve the corresponding constraints (10). If there is no solution, we continue with the next choice. Otherwise let $\sigma^{\vec{p}}$ be the N-memory strategy given by the solution to this equation. If $\mathbb{P}^{\sigma^{\vec{p}}}_q[M[e], \mathsf{Reach}(T_e)] = 0$ for each $e \in E$ and $q \in Q^{\mathsf{no}}_e$, then we return Yes; otherwise we continue with the next choice $(Q^{\mathsf{no}}_e)_{e \in E}$. We return No at the end of if no solution was found.

Let us show that this procedure solves the gap problem. Assume that there exists a strategy τ such that for all environments $e \in E$, we have $\mathbb{P}^{\tau}_{q_0}[M[e], \mathsf{Reach}(T)] \geq \alpha$. By Lemma 24, there exists a N-memory strategy τ' such that in all environments $e \in E$, we have $\mathbb{P}^{\tau'}_{q_0}[M[e], \mathsf{Reach}(T)] \geq \alpha - \varepsilon$. Hence (10) must have a solution corresponding to this strategy for some choice of the sets $(Q^{\mathsf{no}}_e)_{e \in E}$, and the procedure returns Yes. Assume now that no strategy achieves the threshold $\alpha - \varepsilon$. In particular, no N-memory strategy achieves this threshold, and the procedure returns No.

We now analyze the complexity of the procedure. The value N is double exponential in the size of the input, which means that the size of (10) is also double exponential. Polynomial equations can be solved in polynomial space in the size of the equations [5], so here we can solve (10) in double exponential space.

The Gap Problem for Parity We now extend the previous procedure to solve the quantitative parity gap problem based on the following observations. In $M[e]$, any finite-memory strategy σ induces a Markov chain. Then the probability $\mathbb{P}^{\sigma}_{q_0}[M[e], \varphi]$ of satisfying a parity condition φ is equal to the probability of reaching bottom strongly connected components (BSCC) that are winning[3] for φ in the induced Markov chain [3]. But the set of BSCCs only depends on the support of σ, that is, the set of state-action pairs that have positive probability. When considering an MDP under N-memory strategies, the support is the set of tuples (q, i, a, i') such that from combined state (q, i) the strategy has a nonzero probability of picking action a and updating memory to i'.

We proceed as follows. Given MEMDP M, initial state q_0, parity condition φ, and bound N, we enumerate all supports $S \subseteq Q \times [N] \times A \times [N]$. For each support S, let T^S_e be the set of φ-winning BSCCs in $M[e]$ under a strategy with support S. We apply the reachability procedure described above based on (10) for the target sets $(T^S_e)_e$ augmented with the following constraints: for all $(q, i, a, i') \in S$, we add the constraint $p_{q,i}(a, i') > 0$,

[3] Recall that a BSCC is winning for a parity condition if the smallest priority of its states is even.

and for all others $p_{q,i}(a,i') = 0$. If the answer is Yes for some support S, then we return Yes; otherwise we return No.

This solves the gap problem: if there is τ such that for all environments $e \in E$ we have $\mathbb{P}^{\tau}_{q_0}[M[e],\varphi] \geq \alpha$, then by Lemma 24 there exists a N-memory strategy τ' such that for all environments $e \in E$ we have $\mathbb{P}^{\tau'}_{q_0}[M[e],\varphi] \geq \alpha - \varepsilon$. Let S denote the support of the strategy τ', and let T^S_e be the set of winning BSCCs in $M[e]$ under τ. So (10), instantiated for S has a solution, for some choice of the sets $(Q^{\mathsf{no}}_e)_{e\in E}$, and the procedure returns Yes. Assume now that no strategy achieves the threshold $\alpha - \varepsilon$. In particular, there is no N-memory strategy with any support S that achieves this threshold, and the procedure returns No.

There are an exponential number of possibilities for the choice of support. Moreover, given a support S, each set T^S_e can be determined in polynomial time. Overall, the procedure remains in double exponential space.

▶ **Theorem 26.** *The gap problem can be solved in double exponential space for MEMDPs with parity objectives.*

References

1 T. S. Badings, T. D. Simão, M. Suilen, and N. Jansen. Decision-making under uncertainty: beyond probabilities. *Int. J. Softw. Tools Technol. Transf.*, 25(3):375–391, 2023.

2 C. Baier, M. Größer, and N. Bertrand. Probabilistic ω-automata. *J. ACM*, 59(1):1, 2012.

3 C. Baier and J.-P. Katoen. *Principles of Model Checking.* MIT Press, 2008.

4 P. Buchholz and D. Scheftelowitsch. Computation of weighted sums of rewards for concurrent MDPs. *Math. Methods Oper. Res.*, 89(1):1–42, 2019.

5 J. Canny. Some algebraic and geometric computations in PSPACE. In *Proc. of STOC: Symposium on Theory of Computing*, page 460–467. ACM, 1988.

6 K. Chatterjee, M. Chmelík, D. Karkhanis, P. Novotný, and A. Royer. Multiple-environment Markov decision processes: Efficient analysis and applications. In *Proc. of ICAPS: Automated Planning and Scheduling*, pages 48–56. AAAI Press, 2020.

7 K. Chatterjee, L. Doyen, H. Gimbert, and T. A. Henzinger. Randomness for free. *Information and Computation*, 245:3–16, 2017.

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

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

10 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.

11 E. A. Feinberg and A. Shwartz, editors. *Handbook of Markov Decision Processes - Methods and Applications.* Kluwer, 2002.

12 H. Gimbert and Y. Oualhadj. Probabilistic automata on finite words: Decidable and undecidable problems. In *Proc. of ICALP (2)*, LNCS 6199, pages 527–538. Springer, 2010.

13 O. Goldreich. On promise problems (a survey in memory of Shimon Even [1935-2004]). *Manuscript*, 2005.

14 W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

15 K. J. Åström. Optimal control of Markov processes with incomplete state information I. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.

16 O. Madani, S. Hanks, and A. Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artif. Intell.*, 147(1-2):5–34, 2003.

17 D. A. Martin. The determinacy of Blackwell games. *The Journal of Symbolic Logic*, 63(4):1565–1581, 1998.

18 C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of Markov decision processes. *Math. Oper. Res.*, 12(3):441–450, 1987.

19 M. L. Puterman. *Markov decision processes.* John Wiley and Sons, 1994.

20 N. M. van Dijk R. J. Boucherie. *Markov decision processes in practice.* Springer, 2017.

21 J.-F. Raskin and O. Sankur. Multiple-environment Markov decision processes. In *Proc. of FSTTCS: Foundation of Software Technology and Theoretical Computer Science*, volume 29 of *LIPIcs*, pages 531–543. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2014.

22 M. Suilen, M. van der Vegt, and S. Junges. A PSPACE algorithm for almost-sure Rabin objectives in multi-environment MDPs. In *Proc. of CONCUR: Concurrency Theory*, volume 311 of *LIPIcs*, pages 40:1–40:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024.

23 M. van der Vegt, N. Jansen, and S. Junges. Robust almost-sure reachability in multi-environment MDPs. In *Proc. of TACAS: Tools and Algorithms for the Construction and Analysis of Systems*, LNCS 13993, pages 508–526. Springer, 2023.