Deciding the Value 1 Problem for sharp-acyclic Partially Observable Markov Decision Processes

Hugo Gimbert, Youssouf Oualhadj

Deciding the Value 1 Problem for ♯-acyclic Partially Observable Markov Decision Processes

Hugo Gimbert[1] and Youssouf Oualhadj[2]

[1] LaBRI, CNRS, France
hugo.gimbert@labri.fr
[2] LIF, université d'Aix Marseille
youssouf.oualhadj@lif.univ-mrs.fr

Abstract. The value 1 problem is a natural decision problem in algorithmic game theory. For partially observable Markov decision processes with reachability objective, this problem is defined as follows: are there observational strategies that achieve the reachability objective with probability arbitrarily close to 1? This problem was shown undecidable recently. Our contribution is to introduce a class of partially observable Markov decision processes, namely ♯-*acyclic* partially observable Markov decision processes, for which the value 1 problem is decidable. Our algorithm is based on the construction of a two-player perfect information game, called the knowledge game, abstracting the behaviour of a ♯-acyclic partially observable Markov decision process $\mathcal{M}$ such that the first player has a winning strategy in the knowledge game if and only if the value of $\mathcal{M}$ is 1.

1 Introduction

Partially Observable Markov Decision Processes (POMDP for short) Markov decision processes (MDPs) are well established tool for modelling systems that mix both probabilistic and nondeterministic behaviours. The nondeterminism models the choices of the system supervisor (the controller) and the probabilities describe the environment behaviours. When the system offers full information, it is rather easy for the controller to make the best choice, this follows from the fact that fully observable MDPs enjoy good algorithmic properties. For instance ω-regular objectives such as parity objective can be solved in polynomial time [10, 8], as well as quantitative objectives such as average and discounted reward criterions [11, 16]. Moreover, optimal strategies always exist for any tail winning condition [4, 14]. Unfortunately, the assumption that a real life system offers a full observability is not realistic. Indeed, an everyday system cannot be made fully monitored because it is either too large (e.g. information system), or implementing full monitoring is too costly (e.g. subway system), or even not possible (e.g. electronic components of an embedded system). That is why partially observable Markov decision processes are a better suited theoretical tool for modelling real life system. In a POMDP, the state space is partitioned and the decision maker cannot observe the states themselves but only the partition they belong to also called the *observation*. Therefore, two executions that cary the same observations and the same actions are undistinguishable for the controller and hence its choice after both execution is going to be the same. In other words the strategies for the controller are mappings from sequences of observation and actions to actions.

Value 1 Problem This problem is relevant for controller synthesis: given a discrete event system whose evolution is influenced by both random events and controllable actions, it is natural to look for controllers as efficient as possible, i.e. to compute strategies which guarantee a probability to win arbitrarily close 1. This means that the probability of winning could never be 1. This differs from the almost-sure winning problem, where the controller is asked to find a strategy that ensures the objective with probability exactly 1. There are toy examples in which an almost-sure controller does not exist but still there exist controllers arbitrarily efficient, and the system can be considered as safe, see Fig. 1 for example. In this figure, an almost-sure strategy cannot exist since any strategy has to take a risk and guess whether the play has started in the top or bottom part of the game. Nevertheless, one can find a strategy that makes the probability of taking the wrong guess arbitrarily small. Note that in case the example would have been fully observable, the value 1 and the almost-sure winning would coincide, this is actually the case for any tail winning condition for simple stochastic games [14]. This property makes the study of fully observable models way easier and leads in most cases to decidability. But as one can deduce from the same example, almost-sure winning and the value 1 problem do not coincide for POMDPs. Actually, almost-sure winning as well as positive winning for reachability objective are decidable problems [3, 6] as opposed to the value 1 problem.

Related work The value one problem has been left open by Bertoni since the 1970's [1, 2]. Recently, we showed that this problem is undecidable for probabilistic automata [15]. This undecidability result extends to POMDP because they subsume probabilistic automata. Since then, much efforts were put into identifying nontrivial decidable families of probabilistic automata for the value 1 problem. For instance, $\sharp$-acyclic automata [15], structurally simple automata [9], and leaktight automata [12]. The common point between those subclasses is the use of two crucial notions. The first one is the iteration of actions, this operation introduced in [15] for probabilistic automata and inspired by automata-theoretic works, describes the long term effects of a given behaviour. The second one is the limit-reachability. Broadly speaking, limit-reachability, formalises the desired behaviour of a probabilistic automaton that has value 1. Therefore, the technical effort in the previously cited papers consists in relating the operation of iteration with the limit-reachability in a complete and consistent manner. Even though the consistency can be obtained rather easily, the completeness requires restrictions on the model considered. This is not surprising since the general case is not decidable. In this work, we consider POMDP, and identify a subclass for which the value 1 problem is decidable.

Contribution and result we extend the decidability result of [15] to the case of POMDPs. We define a class of POMDPs called $\sharp$-*acyclic* POMDPs and we show that the value 1 problem is decidable for this class.

The techniques we use are new compared to [15]. While in [15] the value 1 problem for $\sharp$-acyclic automata is reduced to a reachability problem in a graph, in the present paper, the value 1 problem for POMDPs is reduced to the computation of a winner in a two-player game. This two-player game is won by the first player if and only if the value of the POMDP is 1. While for $\sharp$-acyclic probabilistic automata the value 1 problem can

be decided in PSPACE, the algorithm for the value 1 problem for $\sharp$-acyclic POMDP runs in exponential time. This algorithm is fix-parameter tractable when the parameter is the number of states per observation.

Even though the class may seem contrived, as the results on probabilistic automata show, this class is useful from a theoretical point of view in the sense that it allows the definition of appropriate formal tools. The main technical challenge was to extend both the notions of iteration and limit-reachability; while in a probabilistic automaton the behaviour of the controller can be described by a finite word, because there is no feed back that the controller could use to change its behaviour. This is not anymore true for a POMDP and behaviour of the controller is described by a (possibly infinite) tree. The choice of the next action actually depends on the sequence of observations received and the actions played. Generalisation from words to trees is in general a nontrivial step and leads both complexity blowups and technical issues. In our case, the effect of this generalisation is mostly notable in the definition of limit-reachability. As one can see in Definition 2 limit-reachability expresses two level of uncertainty as opposed to probabilistic automata where one level is sufficient. The notion of limit-reachability is carefully chosen so that it is transitive in the sense of Lemma 1 and can be algorithmically used thanks to Lemma 3. We believe that this definition can be kept unchanged for handling more general decidable families of POMDPs.

Outline of the paper in Section 2 we introduce POMDPs and related notations. In Section 3 we introduce the class of $\sharp$-acyclic POMDPs and state the decidability of the value 1 problem for $\sharp$-acyclic POMDPS which is our main theorem (c.f. Theorem 2). In Section 4 we define the *knowledge game* and prove the main result.

2 Notations

Given S a finite set, let $\Delta(S)$ denote the set of distributions over S, that is the set of functions $\delta : S \to [0, 1]$ such that $\sum_{s \in S} \delta(s) = 1$. for a distribution $\delta \in \Delta(S)$, the support of δ denoted $\mathsf{Supp}(\delta)$ is the set of states $s \in S$ such that $\delta(s) > 0$. We denote by δ_Q the uniform distribution over a finite set Q.

2.1 Partially Observable Markov Decision Process

Intuitively, to play in a POMDP, the controller receives an observation according to the initial distribution then it chooses an action then it receives an other observation and chooses another action and so on. The goal of the controller is to maximize the probability to reach the set of target states T.

A POMDP is a tuple $\mathcal{M} = (S, A, \mathcal{O}, \mathsf{p}, \delta_0)$ where S is a finite set of states, A is a finite set of actions, $\mathcal{O}$ is a partition of S called the observations, $\mathsf{p} : S \times A \to \Delta(S)$ is a transition function, and δ_0 is an initial distribution in $\Delta(S)$.

We assume that for every state $s \in S$ and every action $a \in A$ the function $\mathsf{p}(s, a)$ is defined, i.e. every action can be played from every state. When the partition described by $O \in \mathcal{O}$ is a singleton $\{s\}$, we refer to state s as observable. An infinite play in a

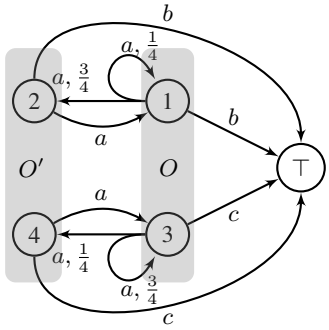


Fig. 1. Partially observable Markov decision process

POMDP is an infinite word in $(\mathcal{O}A)^\omega$, and a finite play is a finite word in $\mathcal{O}(A\mathcal{O})^*$. We denote by Plays the set of finite plays.

Consider the POMDP $\mathcal{M}$ depicted in Fig 1. The initial distribution is at random between states 1 and 3, the play is winning if it reaches $\top$, and the observations are $\mathcal{O} = \{O, O', \{\top\}\}$ where $O = \{1, 3\}$ and $O' = \{2, 4\}$. State $\top$ is observable. The missing transitions lead to a sink and are omitted for the sake of clarity. A play in $\mathcal{M}$ could be $\rho = OaOaO'(aO)^\omega$.

2.2 Outcome and Knowledge

Let $Q \subseteq S$ be a subset and a be a letter, we define $\mathsf{Acc}(Q, a)$ as the set of states $s \in S$ such that there exists $q \in Q$ and $\mathsf{p}(q, a)(s) > 0$.

The outcome of an action a given a subset of states Q is the collection $Q \cdot a$ of states that the controller may believe it is in after it has played action a in one of the states of Q and it has received its observation: $Q \cdot a = \{R \subseteq 2^S \mid \exists O \in \mathcal{O}, R = \mathsf{Acc}(Q, a) \cap O\}$. For a collection of subsets $\mathcal{R} \subseteq 2^S$ we write: $\mathcal{R} \cdot a = \bigcup_{R \in \mathcal{R}} R \cdot a$.

Let $w = O_0 a_1 O_1 \cdots a_n O_n \in \mathsf{Plays}$ be a finite play. The knowledge of the controller after w has occurred is defined inductively as follows:

$$\begin{cases} K(\delta_0, O_0) = \mathsf{Supp}(\delta_0) \cap O_0 \ , \text{ and} \\ K(\delta_0, w) = \mathsf{Acc}(K(\delta_0, O_0 \cdots a_{n-1} O_{n-1}), a_n) \cap O_n \ . \end{cases}$$

It is a an elementary exercise to show that for every strategy σ, the following holds:

$$\mathbb{P}^\sigma_{\delta_0}(\forall n \in \mathbb{N}, S_n \in K(\delta_0, w)) = 1 \ . \tag{1}$$

2.3 Strategies and measure

To play, the controller chooses the next action to apply according to the initial distribution, the sequence of actions played, and the sequence of observations received. Such a way of playing is called *observational*, and any strategy that formalise it is called observational strategy. Formally, an observational strategy for the controller is a function $\sigma : \mathsf{Plays} \to A$. Notice that we consider only pure strategies. This is actually enough since in POMDPs randomised strategies are not more powerful than the pure one [13, 5].

Once an initial distribution δ_0 and a strategy σ are fixed, this defines uniquely a probability measure $\mathbb{P}^\sigma_{\delta_0}$ on $S(AS)^\omega$ as the probability measure of infinite trajectories of the Markov chain whose transition probabilities are fixed by δ_0, σ and $\mathsf{p} : S \times A \to \Delta(S)$. Using the natural projection $\pi : S(AS)^\omega \to \mathcal{O}(A\mathcal{O})^\omega$ we extend the probability measure $\mathbb{P}^\sigma_{\delta_0}$ to $\mathcal{O}(A\mathcal{O})^\omega$.

We define the random variables S_n, A_n, and O_n with values in S, A, and $\mathcal{O}$ respectively that maps an infinite play $w = s_0 a_1 s_1 a_2 s_2 \cdots$ to respectively the n-th state S_n, the n-th action A_n, and the n-th observation $O_n \in \mathcal{O}$ such that $S_n \in O_n$.

2.4 Value 1 problem

The value of a POMPD is the largest probability of winning an objective (reachability in our case) when the play start in the initial distribution. Formally,

Definition 1 (Value). *Let $\mathcal{M}$ be a POMDP, $\delta_0 \in \Delta(S)$ be an initial distribution, and $T \subseteq S$ be a subset of target states, the value of δ_0 in $\mathcal{M}$ is:*

$$\mathsf{Val}_{\mathcal{M}}(\delta_0) = \sup_\sigma \mathbb{P}^\sigma_{\delta_0}(\exists n \in \mathbb{N},\ S_n \in T) \ .$$

The value 1 problem consists in deciding whether $\mathsf{Val}_{\mathcal{M}}(\delta_0) = 1$ for given $\mathcal{M}$ and δ_0.

Example 1. The value of the POMDP of Fig 1 is 1 when the initial distribution is uniform over the set $\{1, 3\}$. Remember that missing edges (for example action c in state 1) go to a losing sink $\bot$, hence the goal of the controller is to determine whether the play is in the upper or the lower part of the game and to play b or c accordingly. Consider the strategy that plays long sequences of a^2 then compares the frequencies of observing O and O'; If O' was observed more than O then with high probability the initial state is 1 and by playing b state $\top$ is reached. Otherwise, with high probability the play is in 3 and by playing c again the play is winning. Note that the controller can make the correct guess with arbitrarily high probability by playing longer sequences of a^2, but it cannot win with probability 1 since it always has to take a risk when choosing between actions b and c. This example shows that the strategies ensuring the value 1 can be quite elaborated: the choice not only depends on the time and the sequence of observations observed, but also depends on the empirical frequency of the observations received.

The value 1 problem is undecidable in general, our goal is to extend the result of [15] and show that the value 1 problem is decidable for $\sharp$*-acyclic POMDP*. The idea is to abstract limit behaviours of finite plays using a finite two-player reachability game on a graph, so that limit-reachability in the POMDP in the sense of Definition 2 coincides with winning the reachability game on the finite graph.

The definition of limit reachability relies on the random variable that gives the probability to be in a set of states $T \subseteq S$ at step $n \in \mathbb{N}$ given the observations received along the $n-1$ previous steps:

$$\phi_n(\delta, \sigma, T) = \mathbb{P}^\sigma_\delta(S_n \in T \mid O_0 A_1 \cdots A_n O_n) \ .$$

Definition 2 (Limit-reachability). *Let $Q \subseteq S$ be a subset of states and $\mathcal{T}$ be a nonempty collection of subsets of states, we say that $\mathcal{T}$ is limit-reachable from S if for every $\varepsilon > 0$ there exists a strategy σ such that:*

$$\mathbb{P}^{\sigma}_{\delta_Q}\left(\exists n \in \mathbb{N},\ \exists T \in \mathcal{T},\ \phi_n(\delta_Q, \sigma, T) \geq 1 - \varepsilon\right) \geq 1 - \varepsilon \ ,$$

where δ_Q is the uniform distribution on Q.

The intuition behind this definition is the following: when $\mathcal{T}$ is limit-reachable from Q, then if the play starts from a state randomly chosen in Q, the controller has observational strategies such that with probability arbitrarily close to 1 it will know someday that the play is in one of the sets $T \in \mathcal{T}$ and which set T it is. Limit-reachability enjoys two nice properties. First the value 1 problem can be rephrased using limit-reachability, second limit-reachability is transitive.

Proposition 1. *Assume that T is observable, i.e.*

$$T = \bigcup_{\substack{O \in \mathcal{O} \\ O \cap T \neq \emptyset}} O \ ,$$

then $\mathsf{Val}_{\mathcal{M}}(\delta_0) = 1$ if and only if $2^T \setminus \emptyset$ is limit-reachable from $\mathsf{Supp}(\delta_0)$.

Proposition 1 does not hold in the case where the set of target states is not observable. However, there is a computable linear time transformation from a POMDP $\mathcal{M}$ to a POMDP $\mathcal{M}'$ with a larger set of states whose set of target states is observable and such that a distribution has value 1 in $\mathcal{M}$ if and only if it has value 1 in $\mathcal{M}'$. Therefore, our decidability result holds whether the target states are observable or not.

Limit-reachability is a transitive relation in the following sense.

Lemma 1 (Limit-reachability is transitive). *Let Q be a subset of states and $\mathcal{R}$ be a nonempty collection of subsets. Assume that $\mathcal{R}$ is limit-reachable from Q and $\mathcal{T}$ a nonempty collection of subsets of states is limit-reachable from every subset $R \in \mathcal{R}$. Then $\mathcal{T}$ is limit-reachable from Q.*

The following lemma shows that the definition of limit-reachability is robust to a change of initial distribution as long as the support of the initial distribution is the same.

Lemma 2. *Let $\delta \in \Delta(S)$ be a distribution, $Q \subseteq S$ its support, $\mathcal{R}$ be a nonempty collection of subsets of states. Assume that for every $\varepsilon > 0$ there exists σ such that:*

$$\mathbb{P}^{\sigma}_{\delta}\left(\exists n \in \mathbb{N},\ \exists R \in \mathcal{R},\ \phi_n(\delta, \sigma, R) \geq 1 - \varepsilon\right) \geq 1 - \varepsilon \ ,$$

then $\mathcal{R}$ is limit-reachable from δ_Q.

The above lemma implies that the decision of the value 1 problem depends only on the support of the initial distribution.

3 The ♯-acyclic Partially Observable Markov Decision Processes

In this section we associate with every POMDP $\mathcal{M}$ a two-player zero-sum game on a graph $\mathcal{G}_{\mathcal{M}}$. The construction of the graph is based on a classical subset construction [7] extended with an iteration operation.

3.1 Iteration of actions

Definition 3 (Stability). *Let $Q \subseteq S$ be a subset of states and $a \in A$ be an action, then Q is a-stable if $Q \subseteq \mathsf{Acc}(Q, a)$.*

Definition 4 (a-recurrence). *Let $Q \subseteq S$ be a subset of states and $a \in A$ be an action such that Q is a-stable. We denote by $\mathcal{M}[Q, a]$ the Markov chain with states Q and probabilities induced by a: the probability to go from a state $s \in Q$ to a state $t \in Q$ is $\mathsf{p}(s, a)(t)$. A state s is said to be a-recurrent if it is recurrent in $\mathcal{M}[S, a]$.*

The key notion in the definition of ♯-acyclic POMDPs is *iteration of actions*. Intuitively, if the controller knows that the play is in Q then either someday it will receive an observation which informs it that the play is no more in Q or otherwise it will have more and more certainty that the play is trapped in the set of recurrent states of a stable subset of Q. Formally,

Definition 5 (Iteration). *Let Q be a subset of states, a be an action such that $Q \in Q \cdot a$ and R be the largest a-stable subset of Q. We define*

$$Q \cdot a^{\sharp} = \begin{cases} \{\{a\text{-recurrent states of } R\}\} \cup (Q \cdot a \setminus \{Q\}) & \text{if } R \text{ is not empty} \\ Q \cdot a \setminus \{Q\} & \text{otherwise .} \end{cases}$$

If $Q \cdot a^{\sharp} = \{Q\}$ then Q is said to be $a^{\sharp}$-stable, equivalently Q is a-stable and all states of Q are a-recurrent.

The action of letters and iterated letters is related to limit-reachability:

Proposition 2. *Let $Q \subseteq S$ and $a \in A$. Assume $Q \subseteq O$ for some $O \in \mathcal{O}$. Then $Q \cdot a$ is limit-reachable from Q. Moreover if $Q \in Q \cdot a$, then $Q \cdot a^{\sharp}$ is also limit-reachable from Q.*

Proof. Let $\varepsilon > 0$ and σ be the strategy that plays only a's. Since $Q \subseteq O$, $\mathbb{P}^{\sigma}_{\delta_Q}(O_0 = 0) = 1$. By definition of the knowledge $K(\delta_Q, O_0) = Q$ thus by definition of $Q \cdot a$,

$$\mathbb{P}^{\sigma}_{\delta_Q}(K(\delta_Q, O_0 a O_1) \in Q \cdot a) = 1 \ ,$$

and according to (1), $\mathbb{P}^{\sigma}_{\delta_Q}(S_1 \in K(\delta_Q, O_0 a O_1) \mid O_0 A_1 O_1) = 1$ thus

$$\mathbb{P}^{\sigma}_{\delta_Q}\left(\phi_1(\delta_Q, \sigma, K(\delta_Q, O_0 a O_1) = 1\right) = 1 \ ,$$

and altogether we get

$$\mathbb{P}^{\sigma}_{\delta_Q}\left(\exists T \in Q \cdot a, \phi_1(\delta_Q, \sigma, T) = 1\right) = 1 \ ,$$

which proves that $Q \cdot a$ is limit-reachable from Q.

Assume that $Q \in Q \cdot a$. By definition of limit-reachability, to prove that $Q \cdot a^\sharp$ is limit-reachable from Q, it is enough to show for every $\varepsilon > 0$,

$$\mathbb{P}^\sigma_{\delta_Q}\left(\exists n \in \mathbb{N},\ \exists T \in Q \cdot a^\sharp,\ \phi_n(\delta_Q, \sigma, T) \geq 1 - \varepsilon\right) \geq 1 - \varepsilon \ . \tag{2}$$

Let R be the (possibly empty) largest a-stable subset of Q, and R' the set of a-recurrent states in R. Let $\mathsf{Stay}^n(O)$ be the event

$$\mathsf{Stay}^n(O) = \{\forall k \leq n, O_k = O\} \ .$$

The strategy σ plays only a's thus $\mathbb{P}^\sigma_{\delta_Q}$ coincides with the probability measure of the Markov chain $\mathcal{M}[S, a]$. Almost-surely the play will stay trapped in the set of a-recurrent states. Thus by definition of R',

$$(R' = \emptyset) \implies \left(\mathbb{P}^\sigma_{\delta_Q}\left(\mathsf{Stay}^n(O)\right) \xrightarrow[n\to\infty]{} 0\right) \tag{3}$$

$$(R' \neq \emptyset) \implies \mathbb{P}^\sigma_{\delta_Q}\left(S_n \in R' \mid \mathsf{Stay}^n(O)\right) \xrightarrow[n\to\infty]{} 1 \ . \tag{4}$$

Now we complete the proof of (2). According to (4) if $R' \neq \emptyset$ there exists $N \in \mathbb{N}$ such that $\mathbb{P}^\sigma_{\delta_Q}\left(S_N \in R' \mid \mathsf{Stay}^N(O)\right) \geq 1 - \varepsilon$, thus

$$(R' \neq \emptyset) \implies \mathbb{P}^\sigma_{\delta_Q}\left(\phi_N(\delta_Q, \sigma, R') \geq 1 - \varepsilon \mid \mathsf{Stay}^N(O)\right) = 1 \ . \tag{5}$$

On the other hand if the play is in $\mathsf{Stay}^n(O)$ and not in $\mathsf{Stay}^{n+1}(O)$ it means the controller receives for the first time at step $n+1$ an observation O_{n+1} which is not O. Since $Q \subseteq O$ it means the play has left Q thus $K(\delta_Q, O_0 a O_1 \cdots O_n) = Q$ and $K(\delta_Q, O_0 a O_1 \cdots O_n a O_{n+1}) = K(\delta_Q, Q a O_{n+1}) \in Q \cdot a \setminus \{Q\}$, thus for every $n \in \mathbb{N}$,

$$\mathbb{P}^\sigma_{\delta_Q}\left(\exists T \in Q \cdot a \setminus \{Q\}, \phi_n(\delta_Q, \sigma, T) = 1 \mid \mathsf{Stay}^n(O) \wedge \neg \mathsf{Stay}^{n+1}(O)\right) = 1. \tag{6}$$

Taking (5) and (6) together with the definition of $Q \cdot a^\sharp$ proves (2). □

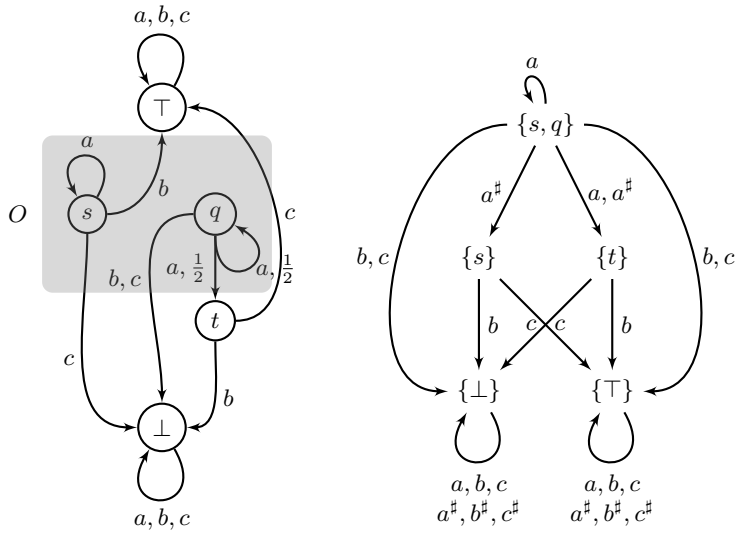
3.2 ♯-acyclic POMDP

The construction of the knowledge graph is based on a classical subset construction extended with the iteration operation.

Definition 6 (Knowledge graph). *Let $\mathcal{M}$ be a POMDP, the knowledge graph $\mathcal{G}_\mathcal{M}$ of $\mathcal{M}$ is the labelled graph obtained as follows:*

- *The states are the nonempty subsets of the observations: $\bigcup_{O \in \mathcal{O}} 2^O \setminus \emptyset$.*
- *The triple (Q, a, T) is an edge if $T \in Q \cdot a$ and the triple $(Q, a^\sharp, T)$ is an edge if $Q \in Q \cdot a$ and $T \in Q \cdot a^\sharp$.*

Example 2. In Fig 2(a) is depicted a POMDP where the initial distribution is at random between states s and q. The states $\top, \bot, t$ are observable and $O = \{s, q\}$. In Fig 2(b) is the knowledge graph associated to it.



(a) A ♯-acyclic POMDP (b) The knowledge graph

Fig. 2. A POMDP and its knowledge graph

Definition 7 (♯-acyclic POMDP). *Let $\mathcal{M}$ be a POMDP and $\mathcal{G}_{\mathcal{M}}$ the associated knowledge graph. $\mathcal{M}$ is ♯-acyclic if the only cycles in $\mathcal{G}_{\mathcal{M}}$ are self loops.*

This condition may seem restrictive, nevertheless, it does not forbid cycles (e.g. [15] for an example).

Of course one can check wether a POMDP is ♯-acyclic or not in exponential time. The main result of the paper is:

Theorem 1. *The value 1 problem is decidable for ♯-acyclic POMDPs. The complexity is polynomial in the size of the knowledge graph, thus exponential in the number of states of the POMDP and fix-parameter tractable with parameter* $\max_{O \in \mathcal{O}} |O|$.

4 Deciding the Value 1

In this section we show that given a POMDP $\mathcal{M}$ and its knowledge graph $\mathcal{G}_{\mathcal{M}}$ there exists a two-player (verifier and falsifier) perfect information game played on $\mathcal{G}_{\mathcal{M}}$ where verifier wins if and only if $\mathsf{Val}_{\mathcal{M}}(\delta_0) = 1$.

4.1 The knowledge game

We first explain how to construct the game and how it is played. Let $\mathcal{M}$ be a POMDP and $\mathcal{G}_{\mathcal{M}}$ be the knowledge graph associated to $\mathcal{M}$. Starting from a vertex Q, the knowledge game is played on $\mathcal{G}_{\mathcal{M}}$ as follows:

- Verifier either chooses an action $a \in A$ or if $Q \in Q \cdot a$ she can also choose an action $a \in A^\sharp$,
- falsifier chooses a successor $R \in S \cdot a$ and $R \in S \cdot a^\sharp$ in the second case,

– the play continues from the new state R.

Verifier wins if the game reaches a subset $R \subseteq T$ of target states.

Definition 8 (♯-reachability). *A nonempty collection of subsets $\mathcal{R}$ is ♯-reachable from a subset Q if there exists a strategy for verifier to reach one of the subsets $R \in \mathcal{R}$ against any strategy of falsifier in the knowledge game.*

Example 3. In the POMDP of Fig 2, assume that the initial distribution δ_0 is at random between state s and q. The value of the initial distribution is 1 because the controller can use the following strategy. Play long sequences of a and if the only observation received is O, with probability arbitrarily close to 1 the play is in state s otherwise with high probability the play would have moved to state q. On the other hand, verifier has a strategy to win from $\{s, q\}$ in the knowledge game. This strategy consists in choosing action $a^\sharp$ from the initial state, then playing action c if falsifier's choice is $\{t\}$ and action b if falsifier's choice is $\{s\}$.

4.2 Proof of Theorem 1

The proof of Theorem 1 is split into Proposition 3 and Proposition 4. The first proposition shows that if verifier has a winning strategy in the knowledge game $\mathcal{G}_\mathcal{M}$, then the value of the POMDP $\mathcal{M}$ is 1. Proposition 3 holds whether the POMDP is ♯-acyclic or not.

Proposition 3. *Let $\mathcal{M}$ be a POMDP with initial distribution δ_0 and let $Q = \mathsf{Supp}(\delta_0)$. Assume that for every observation $O \in \mathcal{O}$ such that $O \cap Q \neq \emptyset$, verifier has a winning strategy in $\mathcal{G}_\mathcal{M}$ from $O \cap Q$. Then $\mathsf{Val}_\mathcal{M}(\delta_0) = 1$.*

Proof. Let $\sigma_\mathcal{M}$ be the winning strategy of the verifier and $\mathcal{T} = 2^T \setminus \emptyset$. The proof is by induction on the maximal number of steps before a play consistent with $\sigma_\mathcal{M}$ reaches $\mathcal{T}$ starting from $Q \cap O$ for all observations O such that $Q \cap O \neq \emptyset$.

If this length is 0 then $\mathsf{Supp}(\delta_0) \subseteq T$ thus $\mathsf{Val}_\mathcal{M}(\delta_0) = 1$.

Otherwise for every observation O such that $Q \cap O \neq \emptyset$, let $a_O = \sigma_\mathcal{M}(Q \cap O)$. Then by induction hypothesis, from every $R \in \mathsf{Supp}(Q \cap O) \cdot a_O$, $\mathsf{Val}_\mathcal{M}(\delta_R) = 1$. Given $\varepsilon > 0$, for every $R \in \mathsf{Supp}((Q \cap O) \cdot a_O)$ let σ_R a strategy in the POMDP to reach T from δ_R with probability at least $1 - \varepsilon$. Let σ be the strategy in the POMDP that receives the first observation O, plays a_O, receives the second observation O_1 then switches to $\sigma_{K(\delta_0, O_0 a_O O_1)}$.

By choice of σ_R, for every state $r \in R$, the strategy σ_R guarantees to reach T from $\delta_{\{r\}}$ with probability at least $1 - |R| \cdot \varepsilon$ thus σ guarantees to reach T from δ_0 with probability at least $1 - |Q| \cdot \varepsilon$. Since this holds for every ε, $\mathsf{Val}_\mathcal{M}(\delta_0) = 1$. □

While it is not too difficult to prove that if verifier wins $\mathcal{G}_\mathcal{M}$ then $\mathsf{Val}_\mathcal{M}(\delta_0) = 1$, the converse is much harder to establish, and holds only for ♯-acyclic POMDPs.

Proposition 4. *Let $\mathcal{M}$ be a ♯-acyclic POMDP and δ_0 be an initial distribution and denote $Q = \mathsf{Supp}(\delta_0)$. Assume that $\mathsf{Val}_\mathcal{M}(\delta_0) = 1$ then for every observation $O \in \mathcal{O}$ such that $O \cap Q \neq \emptyset$, verifier has a winning strategy in $\mathcal{G}_\mathcal{M}$ from $O \cap Q$.*

Lemma 3. *Let Q be a subset such that $Q \subseteq O$ for some observation $O \in \mathcal{O}$. Assume that a nonempty collection of subsets of states $\mathcal{T}$ is limit-reachable from Q, then either $Q \in \mathcal{T}$ for some $T \in \mathcal{T}$ or there exists a a nonempty collection of subsets of states $\mathcal{R}$ such that:*

i) $Q \notin \mathcal{R}$,
ii) $\mathcal{R}$ *is* $\sharp$*-reachable from* Q,
iii) $\mathcal{T}$ *is limit-reachable from every subset in* $\mathcal{R}$.

Proof. If $Q \subseteq T$ for some $T \in \mathcal{T}$, then there is nothing to prove. Assume the contrary. Since $\mathcal{T}$ is limit-reachable from Q, for every $n \in \mathbb{N}$ there exists a strategy σ_n such that:

$$\mathbb{P}^{\sigma_n}_{\delta_Q}\left(\exists m \in \mathbb{N},\ \exists T \in \mathcal{T},\ \phi_m(\delta_Q, \sigma_n, T) \geq 1 - \frac{1}{n}\right) \geq 1 - \frac{1}{n}\ . \tag{7}$$

Let $\pi_n = Oa_1^n Oa_2^n O \cdots$ the unique play consistent with the strategy σ_n such that the observation received all along π_n is O and let $\pi_n^m = Oa_1^{(n)} O \cdots a_m^{(n)} O$. Let $A_Q = \left\{a \in A \mid (Q \in Q \cdot a) \wedge (Q \cdot a^\sharp = \{Q\})\right\}$ and let $d_n = \min\left\{k \mid \sigma_n(\pi_n^k) \notin A_Q\right\}$ with values in $\mathbb{N} \cup \{\infty\}$ and denote $(u_n)_{n\in\mathbb{N}}$ the sequence of words in A^* such that: $u_n = a_1^{(n)} \cdots a_{d_n-1}^{(n)}$.

We need the following preliminary result (proved in Appendix B.2): there exists $\eta > 0$ such that for every $n \geq 0$

$$\mathbb{P}^{\sigma_n}_{\delta_Q}\left(\forall m < d_n,\ \forall T \in \mathcal{T},\ \phi_m(\delta_Q, \sigma_n, T) \leq 1 - \eta\right) = 1\ . \tag{8}$$

As a consequence of (8), it is not possible that for infinitely many n, $d_n = \infty$ otherwise (8) would contradict (7). We assume wlog (simply extract the corresponding subsequence from $(\sigma_n)_n$) that $d_n < \infty$ for every n thus all words u_n and plays $\pi_n^{d_n}$ are finite Since A is finite we also assume wlog that $\sigma_n(\pi_n^{d_n})$ is constant equal to some action $a \in A \setminus A_Q$. Since $a \notin A_Q$ then either $Q \notin Q \cdot a$ or $Q \in Q \cdot a$ and $Q \cdot a^\sharp \neq \{Q\}$. In the first case let $\mathcal{R} = Q \cdot a$ and in the second case let $\mathcal{R} = Q \cdot a^\sharp$.

We show that $\mathcal{R}$ satisfies the constraints of the lemma.

$i)$ holds because $a \notin A_Q$ and by definition of A_Q, $ii)$ holds because either $\mathcal{R} = Q \cdot a$ or $\mathcal{R} = Q \cdot a^\sharp$ hence playing a or $a^\sharp$ is a winning strategy for Verifier.

The proof of $iii)$ is fairly technical and is presented in Appendix B.3 □

Proof (Proposition 4). Let $\mathcal{M}$ be a $\sharp$-acyclic POMDP and δ_0 be an initial distribution. Assume that $\mathsf{Val}(\delta_0) = 1$ then by Proposition 1 we know that $\mathcal{T} = 2^T \setminus \emptyset$ is limit-reachable from $\mathsf{Supp}(\delta_0)$, using the sequence of strategies $(\sigma_n)_{n\in\mathbb{N}}$. Thanks to Lemma 3, we construct a winning strategy for verifier from $\mathsf{Supp}(\delta_0)$: when the current vertex Q is not in $\mathcal{T}$, compute $\mathcal{R}$ given by Lemma 3 and play a strategy to reach one of the vertices $R \in \mathcal{R}$. Because of condition i) of Lemma 3, a play consistent with this strategy will not reach twice in a row the same vertex until the play reaches some vertes $T \in \mathcal{T}$. Since $\mathcal{M}$ is $\sharp$-acyclic, the only loops in $\mathcal{G}_\mathcal{M}$ are self loops and as a consequence the play will necessarilly end up in $\mathcal{T}$. □

Proposition 3 and Proposition 4 lead the following theorem:

Theorem 2. *Given a ♯-acyclic POMDP $\mathcal{M}$ and an initial distribution δ_0. Verifier has a winning strategy in the knowledge game $\mathcal{G}_{\mathcal{M}}$ if and only if $\mathsf{Val}_{\mathcal{M}}(\delta_0) = 1$.*

Theorem 1 follows directly from Theorem 2 and from the fact that the winner of a perfect information reachability game can be computed in quadratic time.

5 Conclusion

We have identified the class of ♯-acyclic POMDP and shown that for this class the value 1 problem is decidable. As a future research, we aim at identifying larger decidable classes such that the answer to the value 1 problem depends quantitatively on the transition probabilities as opposed to ♯-acyclic POMDPs. This would imply an improvement in the definition of the iteration operation, for example considering the stationary distribution of the Markov chain induced by the stable subsets.

References

1. A. Bertoni. The solution of problems relative to probabilistic automata in the frame of the formal languages theory. In *Proc. of the 4th GI Jahrestagung*, volume 26 of *LNCS*, pages 107–112. Springer, 1974.
2. A. Bertoni, G. Mauri, and M. Torelli. Some recursive unsolvable problems relating to isolated cutpoints in probabilistic automata. In *Proceedings of the Fourth Colloquium on Automata, Languages and Programming*, pages 87–94, London, UK, 1977. Springer-Verlag.
3. N. Bertrand, B. Genest, and H. Gimbert. Qualitative determinacy and decidability of stochastic games with signals. In *LICS*, pages 319–328, 2009.
4. K. Chatterjee. Concurrent games with tail objectives. *Theor. Comput. Sci.*, 388(1-3):181–198, 2007.
5. K. Chatterjee, L. Doyen, H. Gimbert, and T. A. Henzinger. Randomness for free. In *MFCS*, pages 246–257, 2010.
6. K. Chatterjee, L. Doyen, and T. A. Henzinger. Qualitative analysis of partially-observable markov decision processes. In *MFCS*, pages 258–269, 2010.
7. K. Chatterjee, L. Doyen, T. A. Henzinger, and J.-F. Raskin. Algorithms for omega-regular games of incomplete information. *LMCS*, 3(3), 2007.
8. K. Chatterjee, M. Jurdziński, and T. A. Henzinger. Quantitative stochastic parity games. SODA '04, pages 121–130, 2004.
9. K. Chatterjee and M. Tracol. Decidable problems for probabilistic automata on infinite words. In *LICS*, pages 185–194, 2012.
10. C. Courcoubetis and M. Yannakakis. The complexity of probabilistic verification. *J. ACM*, 42(4):857–907, 1995.
11. C. Derman. *Finite State Markovian Decision Processes*. Academic Press, Inc., Orlando, FL, USA, 1970.
12. N. Fijalkow, H. Gimbert, and Y. Oualhadj. Deciding the value 1 problem for probabilistic leaktight automata. In *LICS*, pages 295–304, 2012.
13. H. Gimbert. Randomized Strategies are Useless in Markov Decision Processes. July 2009.
14. H. Gimbert and F. Horn. Solving Simple Stochastic Tail Games. SODA '10, page 1000, 01 2010.
15. H. Gimbert and Y. Oualhadj. Probabilistic automata on finite words: Decidable and undecidable problems. In *ICALP*, pages 527–538, 2010.
16. M. L. Putterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, New York, NY, 1994.

Appendix

A Proofs from Section 2

Proposition 5 (Proposition 1 in the paper). *Assume that T is observable, i.e.*

$$T = \bigcup_{\substack{O \in \mathcal{O} \\ O \cap T \neq \emptyset}} O \ ,$$

then $\mathsf{Val}_{\mathcal{M}}(\delta_0) = 1$ *if and only if* $\mathcal{T}$ *is limit-reachable from* $\mathsf{Supp}(\delta_0)$.

Proof. Since T is observable, for every $\varepsilon > 0$,

$$S_n \in T \iff O_n \subseteq T \iff \phi_n(\delta_Q, \sigma, T) = 1 \iff \phi_n(\delta_Q, \sigma, T) \geq 1 - \epsilon \ .$$

As a consequence

$$\begin{aligned} \mathsf{Val}_{\mathcal{M}}(\delta_0) = 1 &\iff \forall \varepsilon > 0,\ \exists \sigma,\ \mathbb{P}^{\sigma}_{\delta_0}(\exists n \in \mathbb{N},\ S_n \in T) \geq 1 - \varepsilon \ , \\ &\iff \forall \varepsilon > 0,\ \exists \sigma,\ \mathbb{P}^{\sigma}_{\delta_0}\left(\mathbb{1}_{O_n \subseteq T}\right) \geq 1 - \varepsilon \ , \\ &\iff \mathbb{P}^{\sigma}_{\delta_0}\left(\exists n \in \mathbb{N},\ \phi_n(\delta_Q, \sigma, T) \geq 1 - \varepsilon\right) \geq 1 - \varepsilon \ . \end{aligned}$$

Where the first equivalence is by definition of the value and the second from the fact that T is observable. □

Lemma 4 (Lemma 1 in the paper). *Let Q be a subset of states and $\mathcal{R}$ be a nonempty collection of subsets. Assume that $\mathcal{R}$ is limit-reachable from Q and $\mathcal{T}$ a nonempty collection of subsets of states is limit-reachable from every subset $R \in \mathcal{R}$. Then $\mathcal{T}$ is limit-reachable from Q.*

Proof. Let $\varepsilon > 0$. Let σ be a strategy such that:

$$\mathbb{P}^{\sigma}_{\delta_Q}\left(\exists n \in \mathbb{N},\ \exists R \in \mathcal{R},\ \phi_n(\delta_Q, \sigma, R) \geq 1 - \frac{\varepsilon}{2}\right) \geq 1 - \frac{\varepsilon}{2} \ ,$$

and for every $R \in \mathcal{R}$ let σ_R such that:

$$\mathbb{P}^{\sigma_R}_{\delta_R}\left(\exists n \in \mathbb{N},\ \exists T \in \mathcal{T},\ \phi_n(\delta_R, \sigma_R, T) \geq 1 - \frac{\varepsilon}{2|R|}\right) \geq 1 - \frac{\varepsilon}{2} \ .$$

Let σ' be the strategy that plays σ until $\phi_n(\delta_Q, \sigma, R) \geq 1 - \frac{\varepsilon}{2}$ for some $R \in \mathcal{R}$, then switches to σ_R. A computation shows that this strategy has the property:

$$\mathbb{P}^{\sigma'}_{\delta_Q}\left(\exists n \in \mathbb{N},\ \exists T \in \mathcal{T},\ \phi_n(\delta_Q, \sigma', T) \geq \left(1 - \frac{\varepsilon}{2}\right) \cdot \left(1 - \frac{\varepsilon}{2}\right)\right) \geq \left(1 - \frac{\varepsilon}{2}\right) \cdot \left(1 - \frac{\varepsilon}{2}\right) \ ,$$

because

$$\left(\phi_n(\delta_R, \sigma_R, T) \geq 1 - \frac{\varepsilon}{2|R|}\right) \implies \forall r \in R, \left(\phi_n(\mathbb{1}_r, \sigma_R, T) \geq 1 - \frac{\varepsilon}{2}\right)$$

□

Lemma 5 (Lemma 2 in the paper). *Let $\delta \in \Delta(S)$ be a distribution, $Q \subseteq S$ its support, $\mathcal{R}$ be a nonempty collection of subsets of states. Assume that for every $\varepsilon > 0$ there exists σ such that:*

$$\mathbb{P}^{\sigma}_{\delta}\left(\exists n \in \mathbb{N},\ \exists R \in \mathcal{R},\ \phi_n(\delta, \sigma, R) \geq 1 - \varepsilon\right) \geq 1 - \varepsilon \ ,$$

then $\mathcal{R}$ is limit-reachable from δ_Q.

Proof. If $\delta = \delta_Q$ then there result is trivial. If not, the result follows from the fact that for every events $E \in s(AS)^{\omega}$, $\varepsilon > 0$, and $n \in \mathbb{N}$:

$$\left(\sum_{s \in Q} \delta(s)\mathbb{P}^{\sigma_n}_{s}(E) \geq 1 - \varepsilon\right) \implies \left(\sum_{s \in Q} \frac{1}{|Q|}\mathbb{P}^{\sigma_n}_{s}(E) \geq 1 - \frac{\varepsilon}{\min_{s \in Q}\{\delta(s)\}}\right) \ .$$

□

The following lemma shows that even though we consider only observable objectives, it is possible to study objectives that are not observable thanks to the following construction.

Lemma 6. *For every POMDP $\mathcal{M}$, there exists a POMDP $\mathcal{M}'$ computable in linear time such that:*

- *the target set in $\mathcal{M}'$ is observable.*
- $\mathsf{Val}_{\mathcal{M}} = 1 \iff \mathsf{Val}_{\mathcal{M}'} = 1$.

Proof. Let $\mathcal{M}$ be a POMDP and let T a set of target states. We construct $\mathcal{M}' = (S', A', \mathcal{O}', \mathsf{p}', \delta'_0)$ such that:

- $S' = (S \times \{0, 1\}) \cup \{\top, \bot\}$.
- $A' = A \cup \{\$\}$ such that for every $s \in Q'$, $\mathsf{p}'((s, 0), \$)(\bot) = 1$ and $\mathsf{p}'((s, 1), \$)(\top) = 1$.
- $\mathsf{p}' : S' \times A' \to \Delta(Q)$ such that for every state $q, t \in S$, action $a \in A$ and $i \in \{0, 1\}$ we have:

$$\mathsf{p}'((s, i), a)(t, 1) = \begin{cases} \mathsf{p}(s, a)(t) \text{ if } (s \in T) \vee (i = 1) \ , \\ 0 \text{ otherwise.} \end{cases}$$

$$\mathsf{p}'((s, i), a)(t, 0) = \begin{cases} \mathsf{p}(s, a)(t) \text{ if } (s \notin T) \wedge (i = 0) \ , \\ 0 \text{ otherwise.} \end{cases}$$

- $\mathcal{O}' = \mathcal{O} \cup \{O_{\top}, O_{\bot}\}$ such that $O_{\top} = \{\top\}$ and $O_{\bot} = \{\bot\}$.
- for every $s \in S$, $\delta'_0(s, 0) = \delta_0(s)$
- $T' = \{\top\}$

We show that $\mathsf{Val}_{\mathcal{M}'} = 1$ if and only if $\mathsf{Val}_{\mathcal{M}} = 1$.

Assume that $\mathsf{Val}_{\mathcal{M}'} = 1$ and let σ' and $\varepsilon > 0$ such that

$$\mathbb{P}^{\sigma'}_{\delta'_0}\left(\exists n \in \mathbb{N}^*,\ S_n = \top\right) \geq 1 - \varepsilon \ ,$$

hence

$$\mathbb{P}^{\sigma'}_{\delta'_0}(\exists n \in \mathbb{N}^*,\ S_{n-1} \in S \times \{1\}) \geq 1 - \varepsilon \ .$$

Let σ be the restriction of σ' on the finite plays defined on $\mathcal{O}(A\mathcal{O})^*$. It follows that:

$$\mathbb{P}^{\sigma}_{\delta_0}(\exists n \in \mathbb{N},\ S_n \in T) \geq 1 - \varepsilon \ .$$

Assume that $\mathsf{Val}_{\mathcal{M}} = 1$ and let σ and $\varepsilon > 0$ such that:

$$\mathbb{P}^{\sigma}_{\delta_0}(\exists n \in \mathbb{N},\ S_n \in T) \geq 1 - \varepsilon \ .$$

Let σ' be a strategy such that for every $\rho \in \mathsf{Plays}$ we have

$$\sigma'(\rho) = \begin{cases} \sigma(\rho) \text{ if } \mathbb{P}^{\sigma}_{\delta_0}(S_n \in Q \times \{1\} \mid \rho) < 1 - \varepsilon \\ \$ \text{ if } \mathbb{P}^{\sigma}_{\delta_0}(S_n \in Q \times \{1\} \mid \rho) \geq 1 - \varepsilon \end{cases}$$

Since by construction of $\mathcal{M}'$ we have

$$\mathbb{P}^{\sigma}_{\delta'_0}(\exists n \in \mathbb{N},\ \forall m \geq n,\ S_m \in Q \times \{1\}) \geq 1 - \varepsilon \ ,$$

it follows that the action $\$$ is chosen at sometime thus

$$\mathbb{P}^{\sigma'}_{\delta'_0}(\exists n \in \mathbb{N},\ S_n = \top) \geq 1 - \varepsilon \ ,$$

which terminates the proof. □

B Proofs from Section 4

B.1 Technical Lemmas

Lemma 7. *Let Q be a subset of states and assume $Q \in Q \cdot a^\sharp$, then $Q \cdot a^\sharp = \{Q\}$.*

Proof. By definition of the iteration operation, Q is the set of a-recurrent states of the largest stable subset of Q. It follows that $Q = Acc(Q, a)$ and all states in Q are a-recurrent thus $Q \cdot a^\sharp = \{Q\}$. □

Lemma 8 (shifting lemma). *Let $f : S^\omega \to \{0, 1\}$ be the indicator function of a measurable event, $\delta \in \Delta(S)$ an initial distribution, and σ a strategy. Then*

$$\mathbb{P}^{\sigma}_{\delta}(f(S_1, S_2, \cdots) = 1 \mid A_1 = a \wedge O_1 = O) = \mathbb{P}^{\sigma'}_{\delta'}(f(S_0, S_1, \cdots) = 1) \ ,$$

where $\forall (s \in S),\ \delta'(s) = \mathbb{P}^{\sigma}_{\delta}(S_1 = s \mid A_1 = a \wedge O_1 = O)$, and $\sigma'(O_2 A_3 \cdots A_n O_n) = \sigma(OaO_2A_3 \cdots A_nO_n)$.

Proof. Using basic definitions, this holds when f is the indicator function of a union of events over S^ω, and the class of events that satisfy this property is a monotone class. □

Lemma 9 (Flooding lemma [15]). *Let $\mathcal{M}$ be a $\sharp$-acyclic POMDP, assume that $\mathcal{O}$ is the signelton $\{S\}$ and for every lettre $a \in A$, $S \cdot a^\sharp = \{S\}$. Then $\{S\}$ is the only limit-reachable collection from S.*

B.2 Proof of Equation (8) of Lemma 3

Let $\mathcal{M}[Q, A_Q, T]$ be the $\sharp$-acyclic automaton with states Q and alphabet A_Q and accepting states T. Almost-surely when playing σ_n from δ_Q all observations are equal to O before step d_n. Thus $\forall T \in \mathcal{T}$ and $m < d_n$,

$$\begin{aligned}\phi_m(\delta_Q, \sigma_n, T) &= \mathbb{P}^{\sigma_n}_{\delta_Q}(S_m \in T \mid O_0 = O_1 = \ldots = O_n = O) \\ &= \mathbb{P}^{\sigma_n}_{\delta_Q}(S_m \in T) = \mathbb{P}_{\mathcal{M}[Q,A_Q,T\cap Q]}(u_n[0,m]) \ , \end{aligned} \tag{9}$$

where $\mathbb{P}_{\mathcal{M}[Q,A_Q,T\cap Q]}(u_n[0,m])$ denotes the probability that the probabilistic automaton $\mathcal{M}[Q, A_Q, T \cap Q]$ accepts the prefix of length m of u_n, denoted $u_n[0,m]$. According to the flooding lemma (Lemma 9) the only subset limit-reachable from Q in the $\sharp$-acyclic automaton $\mathcal{M}[Q, A_Q, T]$ is Q itself. Thus, since forall $T \in \mathcal{T}$, $Q \not\subseteq T$ by hypothesis and by definition of limit-reachability in a probabilistic automaton (see [15])

$$\max_{T\in\mathcal{T}} \sup_{m<d_n} \mathbb{P}_{\mathcal{M}[Q,A_Q,T\cap Q]}(u_n[0,m]) \leq 1-\eta \ ,$$

for some $\eta > 0$ which together with (9) proves (8).

B.3 Proof of point $iii)$ of Lemma 3

We now show that $iii)$ holds, i.e. for every $R \in \mathcal{R}$, the collection $\mathcal{T}$ is limit-reachable from R. According to (8) and (7) for every $n \in \mathbb{N}$ such that $\frac{1}{n} < \eta$,

$$\begin{aligned}&\mathbb{P}^{\sigma_n}_{\delta_Q}\left(\exists m \geq d_n,\ \exists T \in \mathcal{T},\ \phi_m(\delta, \sigma_n, T) \geq 1 - \frac{1}{n}\right) \\ &= \mathbb{P}^{\sigma_n}_{\delta_Q}\left(\exists m \in \mathbb{N},\ \exists T \in \mathcal{T},\ \phi_m(\delta, \sigma_n, T) \geq 1 - \frac{1}{n}\right) \geq 1 - \frac{1}{n} \ .\end{aligned} \tag{10}$$

Let δ' be the distribution $\forall q \in Q,\ \delta'_n(q) = \mathbb{P}^{\sigma'_n}_{\delta_Q}(S_{d_n} = q \mid O_0 = \cdots = O_{d_n} = O)$, and $\forall \pi' \in \mathsf{Plays},\ \sigma'_n(\pi') = \sigma_n(\pi_n^{d_n-1}\sigma_n(\pi_n^{d_n-1})\pi')$. Applying the shifting lemma $d_n - 1$ consecutive times to equation (10), we obtain

$$\mathbb{P}^{\sigma'_n}_{\delta'_n}\left(\exists m \in \mathbb{N}, \exists T \in \mathcal{T},\ \phi_m(\delta'_n, \sigma'_n, T) \geq 1 - \frac{1}{n}\right) \geq 1 - \frac{1}{n} \ .$$

Since all letters played by strategy σ'_n before step d_n are in A_Q then by the flooding lemma again there exist $\eta > 0$ such that $\forall n \in \mathbb{N}, \forall s \in Q, \delta'_n(q) > \eta$. It follows that

$$\mathbb{P}^{\sigma'_n}_{\delta_Q}\left(\exists m \in \mathbb{N}, \exists T \in \mathcal{T},\ \phi_m(\delta_Q, \sigma'_n, T) \geq 1 - \frac{1}{n\cdot\eta}\right) \geq 1 - \frac{1}{n\cdot\eta} \ ,$$

thus we reduced the proof of $iii)$ to the case where forall $n \in \mathbb{N}$, $\sigma_n(O) \notin A_Q$.

Since A_Q is finite we assume from now wlog that there exists $a \in A \setminus A_Q$ such that:

$$(\forall n \in \mathbb{N}, \sigma_n(O) = a) \text{ and } \left(\mathcal{R} = Q \cdot a \text{ or } \mathcal{R} = Q \cdot a^\sharp\right) \ .$$

Assume first that $Q \notin Q \cdot a$ thus $\mathcal{R} = Q \cdot a$. For every $R \in \mathcal{R}$ there is by definition of $Q \cdot a$ some observation $O_R \in \mathcal{O}$ such that $R = Acc(Q, a) \cap O_R$. For every $n \in \mathbb{N}$, let σ_n^R be the strategy defined by $\sigma_n^R(p) = \sigma_n(O \cdot a \cdot p)$. Let $x_R = \mathbb{P}_{\delta_Q}^{\sigma_n}(O_1 = O_R)$ then by definition of $Q \cdot a$ observation O_R may occur with positive probability when playing action a thus $x_R > 0$. Let δ^R the distribution with support R defined by $\delta^R(q) = \mathbb{P}_{\delta_Q}^{\sigma_n}(S_1 = r \mid O_1 = O_R)$. Then

$$\begin{aligned}
&\mathbb{P}_{\delta_Q}^{\sigma_n}\left(\exists m \in \mathbb{N}, \exists T \in \mathcal{T},\ \phi_m(\delta_Q, \sigma_n, T) \geq 1 - \frac{1}{n}\right) \\
&= \mathbb{P}_{\delta_Q}^{\sigma_n}\left(\exists m > 0, \exists T \in \mathcal{T},\ \phi_m(\delta_Q, \sigma_n, T) \geq 1 - \frac{1}{n}\right) \\
&= \sum_{R \in \mathcal{R}} x_R \cdot \mathbb{P}_{\delta^R}^{\sigma_n^R}\left(\exists m \in \mathbb{N}, \exists T \in \mathcal{T},\ \phi_m(\delta^R, \sigma_n^R, T) \geq 1 - \frac{1}{n}\right) \quad ,
\end{aligned}$$

where the first equation holds because by hypothesis there exists no $T \in \mathcal{T}$ such that $Q \subseteq T$ and the second is the shifting lemma. According to (7) the left part of the above equation converges to 1 and since $\forall R \in \mathcal{R}, x_R > 0$ then every subterm of the convex sum in the right part converges to 1 as well. According to Lemma 2, since the support of distribution δ^R is R, it implies that $\mathcal{T}$ is limit-reachable from every support in $\mathcal{R}$. This completes the proof of $iii)$ in the case where $R = Q \cdot a$.

Assume now that $Q \in Q \cdot a$ and $\mathcal{R} = Q \cdot a^\sharp$. Then for every support $R \in (Q \cdot a) \cap (Q \cdot a^\sharp)$ we can use exactly the same proof that in the case where $\mathcal{R} = Q \cdot a$ to show that $\mathcal{T}$ is limit-reachable from R. By definition of $Q \cdot a^\sharp$, the remaining case is the case where R is the set R' of recurrent states of the largest a-stable subset of Q. But since $R' \subseteq Q$, for every $T \in \mathcal{T}$ $\phi_m(\delta_{R'}, \sigma_n, T) \geq \frac{1}{|R'|}\phi_m(\delta_Q, \sigma_n, T)$ and according to Equation (7) it follows that:

$$\mathbb{P}_{\delta_{R'}}^{\sigma_n}\left(\exists m \in \mathbb{N},\ \exists T \in \mathcal{T},\ \phi_m(\delta_{R'}, \sigma_n, T) \geq 1 - \frac{1}{n|R'|}\right) \geq 1 - \frac{1}{n|R'|} \xrightarrow[n \to \infty]{} 1,$$

thus $\mathcal{T}$ is limit-reachable from R'. This completes the proof of $iii)$ in the case where $R = Q \cdot a^\sharp$, and the proof of the lemma.