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 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) is the natural extension of Markov decision processes to the setting of partial information games. In a POMDP, each state is labeled with a color and the decision maker cannot observe the states themselves but only their colors, thus if two plays carry the same colors and the same actions, its choice should be the same in both cases; in other words the strategies for the controller are mappings from sequences of colors and actions to actions.

While in fully observable Markov decision process, ω-regular objectives such as parity game can be solved in polynomial time [7, 5], in POMDPs it is not anymore the case and even deciding whether the value for reachability objectives is 1 or greater than $\frac{1}{2}$ is not decidable [13, 12, 11]. The authors of the present paper proved in a previous paper [11] that this undecidability result holds even if all states are labeled with the same color i.e. for probabilistic automata [13].

The value 1 problem is relevant for controller synthesis: given a discrete event system whose evolution is influenced both by 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 as high as possible. There are toy examples in which an almost-sure controller does not exist but still there exists controllers arbitrarily efficient, and the system can be considered as safe, see Fig. 1 for example.

Partially observable processes are natural models of controllable systems. Although fully observable Markov decision processes are well understood algorithmically and can be solved in polynomial time for most winning conditions [8, 14], they are not very useful as models of controllable systems. Indeed, everyday systems are typically not fully monitored because for example the system is 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).

Related work Previous work has focused on partially observable Markov decision processes with no observation at all, i.e. probabilistic automata. Several subclasses of automata for which the value 1 problem is decidable have been identified, namely ♯-acyclic automata [11], leaktight automata [9], hierarchical automata [2], and structurally simple automata [6].

Another kind of research line has focused on a richer model such as stochastic games with signals [1], but has investigated strategies that achieve a given objectives either almost-surely i.e. with probability 1 or positively i.e. with positive probability.

In this work, we consider one player stochastic game with partial information, and identify a subclass for which the value 1 problem is decidable.

Contribution and result we extend the decidability result of [11] to the case of POMDPs. We define a class of POMDPs called ♯-*acyclic* POMDPs and we show that the value 1 problem is decidable for this class.

The techniques we use are new compared to [11]. While in [11] the value 1 problem for ♯-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: the two-player game is won by the first player if and only if the value of the POMDP is 1. While for ♯-acyclic probabilistic automata the value 1 problem can be decided in PSPACE, the algorithm for the value 1 problem for ♯-acyclic POMDP runs in exponential time. This algorithm is fix-parameter tractable when the parameter is the number of states per observation.

To our opinion, the core notion of this work, is the notion of *limit-reachability*, that we introduced previously [11] for probabilistic automata and which is extended to POMDPs. 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 in the case of POMDP where the behaviour of the controller can be described by a (possibly infinite) tree, in this case the choice of the next action actually depends on the sequence of colors of the states visited. 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 6.

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, namely 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. its goal is to maximise 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 POMDP is an infinite word $w = O_0 a_1 O_1 \cdots \in \mathcal{O}(A\mathcal{O})^\omega$, and a finite play is a finite word in $\mathcal{O}(A\mathcal{O})^*$. We denote by Plays the set of finite plays.

Example 1. 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 possible play in $\mathcal{M}$ is $\rho = OaOaO'(aO)^\omega$.

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:

$$Q \cdot a = \{R \subseteq 2^S \mid \exists O \in \mathcal{O}, R = \mathsf{Acc}(Q, a) \cap O\} \ .$$

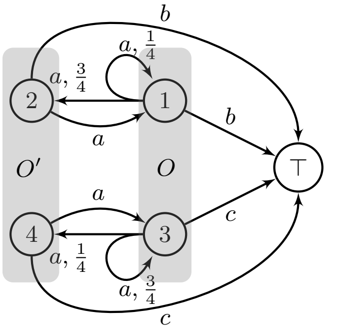


Fig. 1. Partially observable Markov decision process

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 $w = O_0 a_1 O_1 a_2 O_2 \cdots a_n O_n \in \mathsf{Plays}$ be finite play. The knowledge of the controller after w has occurred is defined inductively as follows:

$$K(\delta_0, O_0) = \mathsf{Supp}(\delta_0) \cap O_0$$
$$K(\delta_0, O_0 a_1 O_1 \cdots a_n O_n) = \mathsf{Acc}(K(\delta_0, O_0 a_1 O_1 \cdots a_{n-1} O_{n-1}), a_n) \cap O_n \ .$$

It is a an elementary exercise to show that for every strategy σ,

$$\mathbb{P}^{\sigma}_{\delta_0}(\forall n \in \mathbb{N}, S_n \in K(\delta_0, O_0 A_1 O_1 \cdots A_n O_n)) = 1 \ . \tag{1}$$

2.3 Strategies and measure

To play the controller chooses the next action to apply in function of the initial distribution, the sequence of actions played, and the sequence of observations received along the play. Such strategies are said to be *observational.* 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 randomized strategies are not more powerful than the pure strategies [10, 4].

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 transiton 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 the n-th state, the n-th action, and the n-th observation respectively.

Note that the play presented in Example 1 has measure 0.

Throughout the paper we use standard notation; if X is a real-valued random variable and Y is random variable with finitely or countably many values then $\mathbb{E}[X \mid Y]$ denotes the Y-measurable random variable $w \to \mathbb{E}[X \mid Y = Y(w)]$ and if E is a measurable event then $\mathbb{P}(E \mid Y)$ is the Y-measurable random variable $w \to \mathbb{P}[E \mid Y = Y(w)]$.

2.4 Value 1 problem

In the sequel we will concentrate on reachability objective, hence when referring to the value of a POMDP it is implied that the objective is a reachability objective.

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 2. 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.

The value 1 problem is undecidable in general, our goal is to extend the result of [11] and show that the value 1 problem is decidable for the so called $\sharp$-*acyclic POMDP*. The idea is to represent limit behaviours of sequences of finite plays using a finite two-player 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 past observations:

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

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 $\epsilon > 0$ 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 that when $\mathcal{T}$ is limit-reachable from Q, then when the play starts from a state randomly chosen in Q the controller has strategies so that with probability arbitrarily close to 1 it will know someday according to its observations 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 $\mathcal{T}$ is limit-reachable from $\mathsf{Supp}(\delta_0)$.

The proof is available in the appendix.

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 $\mathcal{M}'$.

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 proof is available in the appendix.

We wil use the following property of limit-reachability. 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:*

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

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

The proof is available in the appendix.

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 [3] 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 $Q \cdot a = \{Q\}$.*

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, i.e. $Q \subseteq \mathsf{Acc}(Q, a)$. 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 $\mathcal{M}[Q, a]$.*

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 $a \in A$. 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. Since a first observation O_0 associated to the initial state is received before the first choice of action of controller, we can assume wlog that this observation is known in advance and assume thus that there exists $O \in \mathcal{O}$ such that $Q \subseteq O$, which makes the proof more readable.

Let $\varepsilon > 0$ and σ the strategy that play sonly a's. By definition of the knowledge $K(\delta_Q, O_0) = Q$ thus by definition of $Q \cdot a$,

$$\mathbb{P}^{\sigma}_{\delta_Q}(K(\delta_Q, O_0aO_1) \in Q \cdot a) = 1 \ ,$$

and according to (1), $\mathbb{P}^{\sigma}_{\delta_Q}(S_1 \in K(\delta_Q, \delta_Q, O_0aO_1) \mid O_0A_1O_1) = 1$ thus

$$\mathbb{P}^{\sigma}_{\delta_Q}\left(\mathbb{P}^{\sigma}_{\delta_Q}\left(\phi_1(\delta_Q, \sigma, K(\delta_Q, O_0aO_1)) = 1\right)\right) = 1 \ ,$$

and altogether we get

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

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

Assume that $Q \in Q \cdot a$ then there exists an observation $O \in \mathcal{O}$ such that $Q \subseteq O$. To prove that $Q \cdot a^\sharp$ is limit-reachable from Q, we are going 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 the (possibly empty) largest a-stable subset of Q, and R' the set of a-recurrent states in R. Let $Stay^n(O)$ the event

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

We show that :

$$\mathbb{P}^{\sigma}_{\delta_Q}\left(S_n \in R' \mid Stay^n(O)\right) \xrightarrow[n\to\infty]{} 1 \ , \tag{3}$$

which is intuitively obvious: if the controller only plays actions a and always receives observation O then chances that the play enters the stable subset R are larger and larger with time. Once in R, the play is a run of the Markov chain induced by R and a thus it will enter the set of recurrent states R' someday. Let $X = Q \setminus R$. We distinguish between the case where $X \neq \emptyset$ and the case where $X = \emptyset$.

If $X \neq \emptyset$, let $x = \min_{s \in X} \sum_{t \notin Q} \mathsf{p}(s, a)(t)$, then by definition of X and R we have $x > 0$, and:

$$\mathbb{P}^{\sigma}_{\delta_Q}\left(\forall k \leq n+1, S_k \in X \mid Stay^n(O)\right) \leq (1-x)^n \xrightarrow[n\to\infty]{} 0 \ .$$

Since $Q \in Q \cdot a$ then $K(\delta_Q, O, a, O, \ldots, a, O) = Q$ and according to (1) the play stays in Q as long as the controller receives only observation O. Since $Q = X \cup R$, it follows that

$$\mathbb{P}^{\sigma}_{\delta_Q}(\exists m \leq n, S_m \in R \mid Stay^n(O)) \xrightarrow[n\to\infty]{} 1 \ . \tag{4}$$

Since R is a-stable,

$$\mathbb{P}^{\sigma}_{\delta_Q}(\exists m \leq n, \forall m \leq k \leq n, S_k \in R \mid Stay^n(O)) \xrightarrow[n\to\infty]{} 1 \ , \tag{5}$$

and since the set of recurrent states R' is reached almost-surely in the finite Markov chain $\mathcal{M}[R, a]$, equation (3) follows (we skip a few uninteresting technical details).

In the case where $X = \emptyset$, Equation (5) applies directly and (3) follows.

Now it is easy to get (2). According to (3) there exists $N \in \mathbb{N}$ such that $\mathbb{P}^{\sigma}_{\delta_Q}\left(S_N \in R' \mid Stay^N(O)\right) \geq 1 - \varepsilon$, thus

$$\mathbb{P}^{\sigma}_{\delta_Q}\left(\phi_N(\delta_Q, \sigma, R') \geq 1 - \varepsilon \mid \mathrm{Stay}^N(O)\right) = 1 \ . \tag{6}$$

On the other hand if the play is $\mathrm{Stay}^n(O)$ and not in $\mathrm{Stay}^{n+1}(O)$ it means the controller receives for the first time at step $n+1$ a signal 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 $n \in \mathbb{N}$,

$$\mathbb{P}^{\sigma}_{\delta_Q}\left(\exists T \in Q \cdot a \setminus \{Q\}, \phi_n(\delta_Q, \sigma, T) = 1 \mid \mathrm{Stay}^n(O) \wedge \neg\mathrm{Stay}^{n+1}(O)\right) = 1. \tag{7}$$

Taking (6) and (7) together with the definition of $Q \cdot a^\sharp$ proves (2). □

3.2 ♯-acyclicPOMDP

The construction of the knowledge graph is based on a classical subset construction [3] 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:*

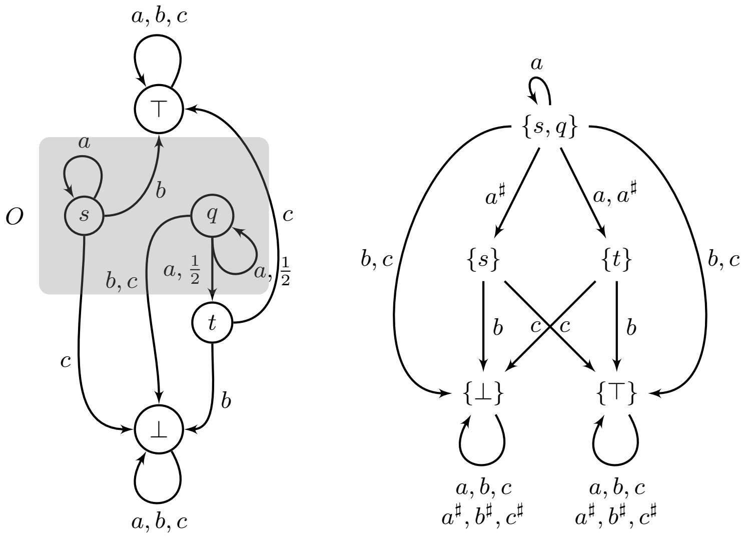
- *The states are the nonempty subsets of the observations: $\bigcup_{O \in \mathcal{O}} 2^O \setminus \emptyset$.*
- *The triple (Q, a, T) is an edge if $\{T\} \in Q \cdot a$ and the triple $(Q, a^\sharp, T)$ is an edge if $\{Q\} \in Q \cdot a$ and $\{T\} \in Q \cdot a^\sharp$.*

Example 3. 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.

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.*

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* $\max_{O \in \mathcal{O}} |O|$.



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

Fig. 2. A POMDP and its knowledge graph

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 $\mathsf{Val}_{\mathcal{M}} = 1$.

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 subset 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 the 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 such that all states in R are 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 a subset $R \in \mathcal{R}$ against any strategy of falsifier in the knowledge game.*

Example 4. 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 play long sequences of a and if the only observation received is O, then 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 $\mathsf{Val}_\mathcal{M} = 1$. Proposition 3 holds whether the POMDP is $\sharp$-acyclic or not.

Proposition 3. *Let $\mathcal{M}$ be a POMDP with initial distribution δ_0 and denote $Q = \mathsf{Supp}(\delta_0)$. 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 of 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, $\mathsf{Val}_\mathcal{M}(\delta_0) = 1$. □

Proposition 4. *Let $\mathcal{M}$ be a $\sharp$-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 $O \cap Q$.*

The proof of this proposition relies on the following several lemmata, the missing proofs are available in the appendix..

Lemma 3. *Let Q be a subset of states and assume $Q \in Q \cdot a^\sharp$, then $Q \cdot a^\sharp = \{Q\}$.*

Lemma 4 (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*

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

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

The following lemma is the key to the decidability of the value 1 problem for $\sharp$-acyclic automata in [11].

Lemma 5 (Flooding lemma [11]). *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.*

The key to the result, and the main technical contribution of this paper is the following lemma.

Lemma 6. *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 $Q \in \mathcal{T}$ 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 $Q \in \mathcal{T}$, then there is nothing to prove. Asume $Q \notin \mathcal{T}$. Since $\mathcal{T}$ is limit-reachable from Q, for every $n \in \mathbb{N}$ there exists a strategy σ_n such that:

$$\mathbb{P}^{\sigma_n}_{\delta_Q}\left(\exists m \in \mathbb{N},\ \exists T \in \mathcal{T},\ \phi_m(\delta_Q, \sigma_n, T) \geq 1 - \frac{1}{n}\right) \geq 1 - \frac{1}{n} \quad . \tag{8}$$

Let $\pi_n = Oa_1^nOa_2^nO\cdots$ the unique play consistent with the strategy σ_n such that the observation received all along π_n is O and $\pi_n^m = Oa_1^nOa_2^nO\cdots a_m^nO$. 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 $\mathbb{N} \cup \{\infty\}$ be the possibly infinite number of steps such that the strategy σ_n plays all its actions in A_Q as long as the observation is always O 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: there exists $\eta > 0$ such that

$$\mathbb{P}^{\sigma_n}_{\delta_Q}\left(\exists m < d_n,\ \exists R \in \mathcal{R},\ \phi_m(\delta_Q, \sigma_n, R) \leq 1 - \eta\right) = 1 \quad . \tag{9}$$

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

$$\begin{aligned}\phi_m(\delta_Q, \sigma_n, R) &= \mathbb{P}^{\sigma_n}_{\delta_Q}(S_m \in R \mid O_0 = O_1 = \ldots = O_n = O) \\ &= \mathbb{P}^{\sigma_n}_{\delta_Q}(S_m \in R) = \mathbb{P}_{\mathcal{M}[Q,A_Q,R]}(u_n[0,m]) \quad ,\end{aligned} \tag{10}$$

where $\mathbb{P}_{\mathcal{M}[Q,A_Q,R]}(u_n[0,m])$ denotes the probability that the probabilistic automaton $\mathcal{M}[Q, A_Q, R]$ accepts the prefix of length m of u_n, denoted $u_n[0,m]$. According to the flooding lemma(Lemma 5) the only subset limit-reachable from Q in the $\sharp$-acyclic automaton $\mathcal{M}[Q, A_Q, R]$ is Q itself. Thus, since $Q \notin \mathcal{R}$ and by definition of limit-reachability in a probabilistic automaton (see [11])

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

for some $\eta > 0$ which together with (10) proves (9).

As a consequence of (9), it is not possible that for infinitely many n, $d_n = \infty$ otherwise (9) would contradict (8). 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 obviously if $a \notin A_Q$ and if $Q \in Q \cdot a$ this is a consequence of Lemma 3.

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.

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 (9) and (8) for every $n \in \mathbb{N}$ such that $\frac{1}{n} < \eta$,

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

Let δ' be the distribution

$$\forall q \in Q,\ \delta'(q) = \mathbb{P}^{\sigma'_n}_{\delta_Q}\left(S_{d_n} = q \mid O_0 = O_1 = \ldots = O_{d_n} = O\right)\ ,$$

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 to this equation $d_n - 1$ consecutive times, we obtain

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

Since all letters played by strategy σ'_n before step d_n are in A_Q then by definition of A_Q the support of δ' is Q. According to Lemma 2, it follows that

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

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 $R = \mathit{Acc}(Q, a) \cap O_R$. 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

$x_R = \mathbb{P}^{\sigma_n}_{\delta_Q}(O_1 = O_R)$ 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 $\delta^R(q) = \mathbb{P}^{\sigma_n}_{\delta_Q}(S_1 = r \mid O_1 = O_R)$. Then

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

According to (8) 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 (8) it follows that:

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

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. □

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 there exists a limit-strategy $(\sigma_n)_{n \in \mathbb{N}}$ from the support $\mathsf{Supp}(\delta_0)$ to $\mathcal{T}$ a nonempty collection of subsets of states. Thanks to Lemma 6, we construct inductively a sequence of collection of support $\mathcal{R}_0, \mathcal{R}_1, \cdots$ such that $\mathcal{R}_0 = \{\mathsf{Supp}(\delta_0)\}$ and for every $i \geq 0$ we have for every $R_i \in \mathcal{R}_i$:

i) $R_i \notin \mathcal{R}_{i+1}$,
ii) $\mathcal{R}_{i+1}$ is $\sharp$-reachable from R_i,
iii) $\mathcal{T}$ is limit-reachable from every support in $\mathcal{R}_{i+1}$.

Since $\mathcal{M}$ is $\sharp$-acyclic, we know that this construction terminates in at most n steps such that $n \leq 2^{2^{|S|}}$ and $\mathcal{R}_n = \mathcal{T}$. Now because $\sharp$-reachability is a transitive relation it follows that $\mathcal{T}$ is $\sharp$-reachable, by definition of $\sharp$-reachability we obtain that there exists a winning strategy for verifier and hence the result. □

Proposition 3 and Proposition 4 lead the following theorem:

Theorem 2. *Given a $\sharp$-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* $\mathsf{Val}(\mathcal{M}) = 1$.

Theorem 1 follows directly from Theorem 2 and from the fact that deciding the winner in a perfect information reachability game is decidable.

5 Conclusion

We have identified the class of $\sharp$-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 $\sharp$-acyclic POMDPs. This would implies 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. Nathalie Bertrand, Blaise Genest, and Hugo Gimbert. Qualitative determinacy and decidability of stochastic games with signals. In *LICS*, pages 319–328, 2009.
2. Rohit Chadha, A. Prasad Sistla, and Mahesh Viswanathan. Power of randomization in automata on infinite strings. In *CONCUR*, pages 229–243, 2009.
3. K. Chatterjee, L. Doyen, T. A. Henzinger, and J.-F. Raskin. Algorithms for omega-regular games of incomplete information. *LMCS*, 3(3), 2007.
4. Krishnendu Chatterjee, Laurent Doyen, Hugo Gimbert, and Thomas A. Henzinger. Randomness for free. In *MFCS*, pages 246–257, 2010.
5. Krishnendu Chatterjee, Marcin Jurdziński, and Thomas A. Henzinger. Quantitative stochastic parity games. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '04, pages 121–130, Philadelphia, PA, USA, 2004. Society for Industrial and Applied Mathematics.
6. Krishnendu Chatterjee and Mathieu Tracol. Decidable problems for probabilistic automata on infinite words. In *LICS*, pages 185–194, 2012.
7. Costas Courcoubetis and Mihalis Yannakakis. The complexity of probabilistic verification. *J. ACM*, 42(4):857–907, 1995.
8. Cyrus Derman. *Finite State Markovian Decision Processes.* Academic Press, Inc., Orlando, FL, USA, 1970.
9. Nathanaël Fijalkow, Hugo Gimbert, and Youssouf Oualhadj. Deciding the value 1 problem for probabilistic leaktight automata. In *LICS*, pages 295–304, 2012.
10. Hugo Gimbert. Randomized Strategies are Useless in Markov Decision Processes. July 2009.
11. Hugo Gimbert and Youssouf Oualhadj. Probabilistic automata on finite words: Decidable and undecidable problems. In *ICALP*, pages 527–538, 2010.
12. Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1-2):5–34, 2003.
13. Azaria Paz. *Introduction to probabilistic automata (Computer science and applied mathematics).* Academic Press, Inc., Orlando, FL, USA, 1971.
14. Martin 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.*

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

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 $\varepsilon > 0$,

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

As a consequence

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

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

Lemma 7 (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 $\epsilon > 0$ Let $Q \not\in \mathcal{R}$ and let $\mathcal{R} \cap \mathcal{T} = \emptyset$ Assume wlog that $(\sigma)_{n \in \mathbb{N}}$ is a limit-strategy from Q to $\mathcal{R}$ and from every $R \in \mathcal{R}$ to $\mathcal{R}$. The result follows from the fact that

$$\mathbb{P}^{\sigma_n}_{\delta_Q}(\exists m' > m > 0,\ \exists (R, T) \in \mathcal{R} \times \mathcal{T},\ \phi_m(\delta_Q, \sigma_n, R) \cdot \phi_{m'}(\delta_R, \sigma_n, T) \geq 1 - 2\varepsilon) \\ \geq 1 - 2\varepsilon \enspace .$$

□

Lemma 8 (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:*

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

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 $n \in \mathbb{N}$:

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

□

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 9. *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

$$\mathbb{P}_{\delta_0'}^{\sigma'}(\exists n \in \mathbb{N}^*,\ S_n = \top) \geq 1 - \varepsilon \ ,$$

hence

$$\mathbb{P}_{\delta_0'}^{\sigma'}(\exists n \in \mathbb{N}^*,\ S_{n-1} \in S \times \{1\}) \geq 1 - \varepsilon \ .$$

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

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

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

Lemma 10 (Lemma 3 of the paper). *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 opeation, 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 11 (Lemma 4 of the paper (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 $\sigma'(O_2 A_3 \cdots A_n O_n) = \sigma(O a O_2 A_3 \cdots A_n O_n)$.

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. □