Limit-sure Reachability for Small Memory Policies in POMDPs is NP-complete

Ali Asadi[1] **Krishnendu Chatterjee**[1] **Raimundo Saona**[1] **Ali Shafiee**[1]

[1]Insitute of Science and Technology Austria, Klosterneuburg, Austria

Abstract

A standard model that arises in several applications in sequential decision-making is partially observable Markov decision processes (POMDPs) where a decision-making agent interacts with an uncertain environment. A basic objective in POMDPs is the reachability objective, where, given a target set of states, the goal is to eventually arrive at one of them. The limit-sure problem asks whether reachability can be ensured with probability arbitrarily close to 1. In general, the limit-sure reachability problem for POMDPs is undecidable. However, in many practical cases, the most relevant question is the existence of policies with a small amount of memory. In this work, we study the limit-sure reachability problem for POMDPs with a fixed amount of memory. We establish that the computational complexity of the problem is NP-complete.

1 INTRODUCTION

MDPs and POMDPs. A standard model in sequential decision-making is Markov decision processes (MDPs) [Bellman, 1957, Howard, 1960], which represents dynamical systems with both nondeterministic and probabilistic behavior. MDPs provide the framework to model and solve control and probabilistic planning and decision-making problems [Filar and Vrieze, 1997, Puterman, 2014] where the nondeterminism represents the choice of the control actions for the controller (or agent) and the probabilistic behavior represents the stochastic response of the system to control actions. In perfectly observable MDPs the controller observes the evolution of the states of the system precisely, whereas in partially observable MDPs (POMDPs) the state space is partitioned according to observations for the controller, i.e., the controller can only view the observation of the current state (the partition the state belongs to) and not the precise state [Bertsekas, 1976, Papadimitriou and Tsitsiklis, 1987]. POMDPs are widely used in several applications, including computational biology [Durbin et al., 1998], speech processing [Mohri, 1997], image processing [Culik and Kari, 1997], software verification [Černý et al., 2011], robot planning [Kress-Gazit et al., 2009, Kaelbling et al., 1998], and reinforcement learning [Kaelbling et al., 1996].

Reachability objectives and computational problems. A basic and fundamental objective in POMDPs is the reachability objective. Given a set of target states, the reachability objective requires that some target state is visited at least once. A policy is a recipe that resolves the choice of control actions. The main computational problems for POMDPs with reachability objectives are: (a) the quantitative problem asks if, for a fixed $\lambda \in (0, 1)$, there exists a policy that ensures the reachability objective with probability at least λ; and (b) the qualitative problem has two variants: (i) almost-sure winning problem asks if there exists a policy that ensures the reachability objective with probability 1; and (ii) limit-sure winning problem asks whether, for every $\lambda < 1$, there exists a policy that ensures the reachability objective with probability at least λ (i.e., ensuring the reachability objectives with probability arbitrarily close to 1).

Significance of qualitative problems. The qualitative problem of limit-sure winning is of great significance in several applications. For example, in the analysis of randomized embedded schedulers [Baruah et al., 1992, Chatterjee et al., 2013], the important question is whether every thread progresses with probability arbitrarily close to 1. Moreover, in applications where it might be sufficient that the correct behavior happens with probability at least $\lambda < 1$, the correct choice of the threshold λ can still be challenging, due to simplifications and imprecisions introduced during modeling. In cases where almost-sure winning cannot be ensured, limit-sure winning provides the strongest guarantee as compared to quantitative problems. Besides its importance in practical applications, almost-sure and limit-sure convergence, like convergence in expectation, is a fundamental

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:238–256.

Table 1: Complexity of quantitative, almost-sure, and limit-sure problems for general and constant-memory policies. Our contribution is marked in bold.

Problem	Policies	
	General	Constant memory
Almost-sure	EXPTIME	NP-complete
Limit-sure	Undecidable	**NP-complete** Theorem 1, Corollary 2
Quantitative	Undecidable	ETR-complete

concept in probability theory, and provides the strongest probabilistic guarantees [Durrett, 2019].

Previous results. The quantitative analysis problem for POMDPs with reachability objectives is undecidable [Paz and Rheinboldt, 1971], and the undecidability result even holds for any approximation [Madani et al., 2003]. In contrast, the complexities of the qualitative analysis problems are as follows: (a) the almost-sure winning problem is EXPTIME-complete [Chatterjee et al., 2010, Baier et al., 2012]; and (b) the limit-sure winning problem is undecidable [Gimbert and Oualhadj, 2010, Chatterjee and Henzinger, 2010].

Small-memory policies. While the computational complexities for the general problems are very high (undecidable in several cases), the same computational questions restricted to policies with small or constant amount of memory are important. This is an interesting theoretical question and is practically relevant as the existence of a small-sized controller is desirable in all applications. The existence of small-memory policies for almost-sure winning was studied by [Chatterjee et al., 2016], and proved to be NP-complete. However, the quantitative problem is ETR-complete [Junges et al., 2018, 2021], even for memoryless policies, where ETR is the existential theory of the reals, and it is a major open question if ETR is in NP or not. The complexity of the limit-sure problem with respect to small-memory policies, which reduces to memoryless policies, has not been studied and is the focus of this work.

Our contributions. In contrast to perfect-observation MDPs where almost-sure winning coincides with limit-sure winning, we show that, in POMDPs, almost-sure winning is different from limit-sure winning, see Example 1. Our main contribution to the limit-sure winning problem for POMDPs with reachability objectives with respect to constant memory policies is to establish that the computational complexity is NP-complete. Table 1 summarizes the complexity results.

Technical contributions. While we establish the same computational complexity of NP-completeness for limit-sure winning as for almost-sure winning, there are significant technical challenges. For example, in general, the almost-sure problem is EXPTIME-complete whereas the limit-sure problem is undecidable, which highlights that they are different problems. Under memoryless policies, if a policy is almost-sure winning, then playing the actions in its support uniformly at random is also almost-sure winning, so it suffices to guess the support of actions. In contrast, we show that memoryless policies that are witness of the limit-sure winning property are more refined: first, they are functional policies; second, there is a notion of ranks over actions where rank k actions are played with probability proportional to ε^k.

Related works. The area of POMDPs with applications is a huge and active research area. POMDPs with reachability objectives have been considered in the probabilistic automata theory community [Gimbert and Oualhadj, 2010, Chatterjee and Henzinger, 2010, Chatterjee et al., 2010, Baier et al., 2012] as well as in the probabilistic planning community [Kress-Gazit et al., 2009, Kaelbling et al., 1998].

The contingent or strong planning considers probabilistic choices as an adversary and is different from the qualitative winning problems we consider. The strong cyclic planning problem is EXPTIME-complete [Kaelbling et al., 1998] and is closer to the almost-sure winning problem, but there are subtle differences, see [Chatterjee et al., 2016].

The almost-sure winning problem is considerably different from limit-sure winning which is in general undecidable, and none of the previous approaches apply to the limit-sure winning problem under small-memory policies.

Similarly, the connection between small-memory policies for POMDPs and parametric Markov chains (pMCs) was established by [Junges et al., 2018] for quantitative reach-avoid objectives. Later, [Junges et al., 2021] proved that some qualitative reachability problems for pMCs are all in NP. They qualitative problems they considered include almost-sure reachability but not limit-sure reachability. Therefore, this line of work doe not apply to the limit-sure winning problem under small-memory policies either.

Detailed proofs omitted due to lack of space are presented in the Appendix.

2 PRELIMINARIES

Notation. For a positive integer n the set $\{1, 2, \dots, n\}$ is denoted by $[n]$. For a set $\mathcal{A}$, the set of probability distributions over $\mathcal{A}$ is denoted by $\Delta(\mathcal{A})$. The probability distribution that assigns probability one to an element $a \in \mathcal{A}$ is denoted by $\mathbb{1}[a]$. The disjoint union of sets is denoted by $\sqcup$. For convenience, we will exchange the roles of λ and $1 - \varepsilon$ depending on the context.

POMDPs. A Partially Observable Markov Decision Process (POMDP) is a tuple $P = (\mathcal{S}, \mathcal{A}, \delta, \mathcal{Z}, o, s_0)$ where:

- $\mathcal{S}$ is a finite set of states;
- $\mathcal{A}$ is a finite set of actions;
- $\delta\colon \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is a probabilistic transition function that, given a state s and an action a, returns the probability distribution over the successor states, i.e., the transition probability from s to s' given a is denoted by $\delta(s,a)(s')$;
- $\mathcal{Z}$ is a finite set of observations;
- $o\colon \mathcal{S} \to \mathcal{Z}$ is an observation function that maps every state to an observation which, for simplicity, and without loss of generality, we consider that o is a deterministic function [Chatterjee et al., 2015, Remark 1];
- $s_o \in \mathcal{S}$ is the unique initial state.

If $|\mathcal{Z}| = 1$, then we call the POMDP a blind MDP since the controller receives no information through the observations. In this case, we identify the blind MDP with the tuple $(\mathcal{S}, \mathcal{A}, \delta, s_0)$. Similarly, if $|\mathcal{A}| = |\mathcal{Z}| = 1$, then we call the POMDP a Markov chain, which coincides with the classic definition, and identify it simply with the tuple $(\mathcal{S}, \delta, s_0)$.

Plays and cones. A play (or a path) in a POMDP is an infinite sequence $(s_0, a_0, s_1, a_1, \ldots)$ of states and actions such that, for all $i \geq 0$, we have $\delta(s_i, a_i)(s_{i+1}) > 0$. For a finite prefix $\omega \in (\mathcal{S} \times \mathcal{A})^*$ of a play, the cone given by ω is the set of plays with ω as the prefix, and the last state of ω, or s_0 if ω is empty, is denoted by Last(ω). For a finite prefix $\omega = (s_0, a_0, s_1, a_1, \ldots, s_n, a_n)$ the sequence of observations and actions associated with ω is denoted by $o(\omega) = (o(s_0), a_0, o(s_1), a_1, \ldots, o(s_n), a_n) \in (\mathcal{Z} \times \mathcal{A})^*$, which we call an observable history.

Policies. A policy is a recipe to extend prefixes of plays and is a function $\sigma\colon \mathcal{Z} \times (\mathcal{A} \times \mathcal{Z})^* \to \Delta(\mathcal{A})$ that, given a finite observable history, selects a probability distribution over the actions. The set of all policies is denoted by Σ.

Policy with memory. A policy with memory is a tuple $\sigma = (\sigma_a, \sigma_u, \mathcal{M}, m_0)$ where: (i) $\mathcal{M}$ is a finite set of memory states; (ii) the function $\sigma_a\colon \mathcal{M} \times \mathcal{Z} \to \Delta(\mathcal{A})$ is the action selection function that, given the current memory state and observation, gives the probability distribution over actions; (iii) the function $\sigma_u\colon \mathcal{M}\times\mathcal{Z}\times\mathcal{A} \to \mathcal{M}$ is the memory update function that, given the current memory state, observation, and action, updates the memory state; and (iv) the memory state $m_0 \in \mathcal{M}$ is the initial memory state. The set of all policies with memory amount m is denoted by Σ_m.

Memoryless policies. A policy σ is memoryless (or observation-stationary) if it depends only on the current observation, i.e., for every two histories ω and ω', if $o(\text{Last}(\omega)) = o(\text{Last}(\omega'))$, then $\sigma(o(\omega)) = \sigma(o(\omega'))$. Therefore, a memoryless policy is just a mapping from observations to a distribution over actions $\sigma\colon \mathcal{Z} \to \Delta(\mathcal{A})$. The set of all memoryless policies corresponds to Σ_1.

Probability measure. Given a policy σ and a starting state s_0, the unique probability measure over Borel sets of infinite plays obtained given σ is denoted by $\mathbb{P}^{\sigma}_{s_0}(\cdot)$, which is defined by Carathéodory's extension theorem by extending the natural definition over cones of plays [Billingsley, 2012].

Reachability objective and value. Given a set of target states, the reachability objective requires that a target state is visited at least once. For simplicity and w.l.o.g., we consider that there is a single target state $\top \in \mathcal{S}$ since we can always add an additional state with transitions from all target states. Formally, given a target state $\top \in \mathcal{S}$, the reachability objective is $\text{Reach}(\top) = \{(s_i, a_i)_{i\geq 0} \in (\mathcal{S} \times \mathcal{A})^{\mathbb{N}} \mid \exists i \geq 0 : s_i = \top\}$. The reachability value under Σ is $\sup_{\sigma\in\Sigma} \mathbb{P}^{\sigma}_{s_0}(\text{Reach}(\top))$.

Almost-sure winning. A POMDP P with reachability objective Reach$(\top)$ is almost-sure winning under Σ if there is a fixed policy $\sigma \in \Sigma$ such that

$$\mathbb{P}^{\sigma}_{s_0}(\text{Reach}(\top)) = 1\,.$$

Limit-sure winning. A POMDP P with reachability objective Reach$(\top)$ is limit-sure winning under Σ if its reachability value under Σ is 1, i.e., if, for all $\varepsilon > 0$, there is a policy $\sigma_\varepsilon \in \Sigma$ such that $\mathbb{P}^{\sigma_\varepsilon}_{s_0}(\text{Reach}(\top)) \geq 1 - \varepsilon$, or equivalently, if

$$\sup_{\sigma\in\Sigma} \mathbb{P}^{\sigma}_{s_0}(\text{Reach}(\top)) = 1\,.$$

Problems under constant memory. The limit-sure (resp. almost-sure) problem under constant amount of memory $m \geq 1$ asks whether a POMDP P is limit-sure (resp. almost-sure) winning under policies restricted to Σ_m.

3 COMPUTATIONAL COMPLEXITY

In this section, we present the main complexity result and show that almost-sure winning and limit-sure winning are different properties in POMDPs through the following example.

Example 1 (Almost-sure $\neq$ Limit-sure). *Consider a blind MDP (with no helpful observation) with four states and two actions* wait, w, *and* commit, c. *The transitions are such that, under action* wait, *the initial state, s_0, may loop with positive probability or may transition to a second state, s_1, with positive probability. Under action* commit, *the initial state moves to an absorbing state, $\perp$, while the second state reaches the target, $\top$. See Figure 1 for an illustration. The reachability value of this blind MDP is* 1. *On the one hand, there is no policy that guarantees reachability value one,*

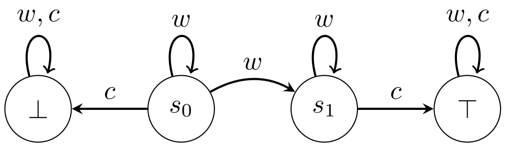

Figure 1: Example of POMDP that is limit-sure winning but not almost-sure winning. Edges represent a positive probability transition between states when the corresponding action in its label is used.

and therefore the blind MDP is not almost-sure winning. On the other hand, for every $\varepsilon > 0$, a policy guaranteeing a reachability probability of at least $1 - \varepsilon$ requires playing action wait *sufficiently many times before playing action* commit. *For every $\varepsilon > 0$, this behavior can be simulated by a distribution over actions that assigns little probability to action* commit. *Therefore, the blind MDP is limit-sure winning, even under memoryless policies.* □

Main novelty of limit-sure vs almost-sure winning. The limit-sure winning property relates to a sequence of policies, as opposed to the almost-sure winning property which relates to a single policy. Moreover, given a sequence of policies $(\sigma_\varepsilon)_{\varepsilon>0}$ that prove the limit-sure winning property, if it exists, the limit policy $\lim_{\varepsilon\to 0^+} \sigma_\varepsilon$ is not a witness of the limit-sure winning property. This is the case in Example 1 where the limit policy applies action *wait* always and does not indicate that the POMDP is limit-sure winning. To preserve the asymptotic information, we work with symbolic or functional policies, called *rank policies*, which assign probabilities to actions based on ranks. For intuition, lower ranks have higher priority, and, if low-rank actions form a cycle, then higher-rank actions determine the exit distribution. In Example 1, a rank policy giving a low rank to action *wait* and a high rank to *commit* reflects that the POMDP is limit-sure winning. Note that considering rank policies with only one rank is equivalent to classic policies that choose actions uniformly at random in its support, which is enough to solve the almost-sure problem [Chatterjee et al., 2016].

We now state the main complexity result.

Theorem 1. *The problem of determining whether a POMDP P with reachability objective is limit-sure winning under memoryless policies is NP-complete.*

The rest of this section is dedicated to the proof of Theorem 1. First, we recall some fundamental concepts. Second, we show the NP upper bound by proving the existence of rank policy witnesses of polynomial size and providing a polynomial-time verifier. Third, we show the NP-hardness by a reduction from 3-SAT. Finally, we present extensions of Theorem 1 for small-memory policies and objectives other than reachability.

3.1 PREVIOUS CONCEPTS

We introduce the most important previous concepts used in our proof.

Real-closed fields. A real-closed field R is a field, i.e., a set on which addition, subtraction, multiplication, and division work as usual, and moreover the intermediate value theorem applies. For an introduction, see [Basu et al., 2006].

Puiseux functions. The set of Puiseux functions is the set of functions $f\colon [0, \varepsilon_0) \to \mathbb{R}$ of the form $f(\varepsilon) = \sum_{i\geq k} c_i \varepsilon^{i/q}$ where $k \in \mathbb{Z}$, i ranges in $\mathbb{Z}$, $c_i \in \mathbb{R}$, $q \in \mathbb{N}$, and $\varepsilon_0 > 0$. The field of Puiseux functions is an important example of a nonarchimedean real-closed field.

Theorem 2 ([Bewley and Kohlberg, 1976, Section 10])**.** *The field of Puiseux functions is real-closed.*

First-order theory of the reals. A sentence in the first-order theory of the reals ϕ is given by

$$Q_1 x_1 Q_2 x_2 \dots Q_k x_k \quad F(x_1, x_2 \dots, x_k)\,,$$

where $Q_i \in \{\exists, \forall\}$ are quantifiers and $F(x_1, x_2 \dots, x_k)$ is a quantifier-free formula in the language of ordered fields with coefficients in a real-closed field. The decision problem for the first-order theory of the reals is, given a sentence ϕ, to decide whether it is true or false. A fundamental result in logic is the following.

Theorem 3 (Tarski-Seidenberg principle [Basu et al., 2006, Theorem 2.80, page 70])**.** *Suppose that R is a real-closed field that contains the real-closed field $\mathbb{R}$. If ϕ is a sentence in the language of ordered fields with coefficients in $\mathbb{R}$, then it is true in $\mathbb{R}$ if and only if it is true in R.*

The following result is a characterization of the reachability value in Markov chains.

Theorem 4 ([Baier and Katoen, 2008, Theorem 10.15, page 762])**.** *Consider a Markov chain with a set of states $\mathcal{S}$ and a target set $\{\top\}$. Then, the reachability value as a function of the initial state is a solution $v^* \in [0,1]^{\mathcal{S}}$ of the system of equations given by $v(\top) = 1$ and, for all $s \in \mathcal{S} \setminus \{\top\}$,*

$$v(s) = \sum_{\tilde{s}\in\mathcal{S}} \delta(s, \tilde{s}) v(\tilde{s})\,,$$

such that, for all other solutions u^*, we have that $v^*(s) \leq u^*(s)$, for all $s \in \mathcal{S}$.

Since we consider Markov chains whose transition probabilities are Puiseux functions, which we call Puiseux Markov chains, we introduce a few concepts from Solan [2003].

Puiseux Markov chains. A Puiseux Markov chain is a family of Markov chains parameterized by ε where the transition function is a Puiseux function $\varepsilon \mapsto \delta^{\varepsilon} \colon \mathcal{S} \to \Delta(\mathcal{S})$. In particular, for each ε, the transition δ^{ε} and the starting state s induces a probability measure $\mathbb{P}_s^{\varepsilon}$.

Reach and exit times, and exit event. For a Markov chain, a state $s \in \mathcal{S}$ and a set of states $\mathcal{B} \subseteq \mathcal{S}$, we consider the following random variables:

$$\begin{aligned} \text{exit}(\mathcal{B}) &\coloneqq \min\{n \geq 0 : s_n \notin \mathcal{B}\}\,, \\ \text{reach}(s) &\coloneqq \min\{n \geq 0 : s_n = s\}\,, \text{ and} \\ \text{Exit}(\mathcal{B}, s) &\coloneqq \{\text{exit}(\mathcal{B}) < \infty\} \cap \{s_{\text{exit}(\mathcal{B})} = s\}\,, \end{aligned}$$

i.e., exit$(\mathcal{B})$ is the first time a state outside of $\mathcal{B}$ is visited, reach(s) is the first time the state s is visited, and Exit$(\mathcal{B}, s)$ is the event of exiting the set $\mathcal{B}$ by visiting state s. In particular, the event Reach$(\tilde{s})$ is equivalent to reach$(\tilde{s}) < \infty$.

The following definition generalizes the concept of communicating class in Markov chains for Puiseux Markov chains.

Communicating class in Puiseux Markov chains. Given a Puiseux Markov chain, a set of states $\mathcal{B} \subseteq \mathcal{S}$ is a communicating class if, for all $s, \tilde{s} \in \mathcal{B}$, we have

$$\begin{aligned} \lim_{\varepsilon \to 0^+} \mathbb{P}_s^{\varepsilon}\big(\text{exit}(\mathcal{B}) < \text{reach}(\tilde{s})\big) = 0\,, \\ \lim_{\varepsilon \to 0^+} \mathbb{P}_s^{\varepsilon}\big(\text{Reach}(\tilde{s})\big) = 1\,, \end{aligned}$$

i.e., starting from s, state $\tilde{s}$ is visited before exiting $\mathcal{B}$. Note that the second condition corresponds to the case of exit$(\mathcal{B})$ = reach$(\tilde{s}) = \infty$ in [Solan, 2003], which is implicitly included in this previous work and prevents that a communicating class consists of unconnected states.

The following concept is key to characterizing events in Puiseux Markov chains.

Exit graph. Given a Puiseux Markov chain and a set of states $\mathcal{B} \subseteq \mathcal{S}$, an exit graph of $\mathcal{B}$, denoted by g, is a directed acyclic graph with edges $E(g) \subseteq \mathcal{B} \times \mathcal{S}$ such that, for all $s \in \mathcal{B}$, there exists $\tilde{s} \in \mathcal{S}$ such that $(s, \tilde{s}) \in E(g)$. The set of all exit graphs of $\mathcal{B}$ is denoted by $G_{\mathcal{B}}$, and all those in which s can reach $\tilde{s}$ by $G_{\mathcal{B}}(s \to \tilde{s})$. The probability of an exit graph g is the product of the probability of each of its transitions defined as $\delta(g) \coloneqq \prod_{(s,\tilde{s}) \in g} \delta(s)(\tilde{s})$. The *weight* of an exit graph g is the leading power in the Puiseux series expansion of the product of the involved transitions defined as

$$w(g) \coloneqq \inf\left\{ r \geq 0 : \lim_{\varepsilon \to 0^+} \frac{\delta^{\varepsilon}(g)}{\varepsilon^r} \neq 0 \right\}\,.$$

The set of exit graphs of $\mathcal{B}$ that have minimal weight is denoted by $G_{\mathcal{B}}^{\min}$, and $G_{\mathcal{B}}^{\min}(s \to \tilde{s})$ for those in which s can reach $\tilde{s}$.

The following result shows that the exit distribution of a communicating class is independent of the initial state within the communicating class.

Theorem 5 ([Solan, 2003, Lemma 3, page 270])**.** *Consider a Puiseux Markov chain and a communicating class $\mathcal{B} \subseteq \mathcal{S}$. Then, the following expression is independent of the initial state $s \in \mathcal{B}$*

$$\delta(\mathcal{B}, \tilde{s}) \coloneqq \lim_{\varepsilon \to 0^+} \mathbb{P}_s^{\varepsilon}(\mathit{Exit}(\mathcal{B}, \tilde{s}))\,.$$

Finally, the following result characterizes the exit "distribution" in terms of exit graphs.

Theorem 6 ([Solan, 2003, Equation 6, page 268])**.** *Consider a Puiseux Markov chain and a communicating class $\mathcal{B} \subseteq \mathcal{S}$. Then, for all $s \in \mathcal{B}$ and $\tilde{s} \in \mathcal{S} \setminus \mathcal{B}$,*

$$\delta(\mathcal{B}, \tilde{s}) = \lim_{\varepsilon \to 0^+} \frac{\sum_{g \in G_{\mathcal{B}}^{\min}(s \to \tilde{s})} \delta^{\varepsilon}(g)}{\sum_{g \in G_{\mathcal{B}}^{\min}} \delta^{\varepsilon}(g)}\,,$$

where the sum over an empty set is 0 *and the quotient* $0/0$ *is also* 0.

We call $\delta(\mathcal{B}, \cdot)$ an exit "distribution" even when it can be constant to zero. Its support corresponds to all the states mapped to a strictly positive value. The following concept allows us to characterize limit-sure reachability in Puiseux Markov chains.

Absorbing communicating class. Given a Puiseux Markov chain, a communicating class $\mathcal{B} \subseteq \mathcal{S}$ is absorbing if its exit distribution has empty support, i.e., $\operatorname{supp}(\delta(\mathcal{B}, \cdot)) = \emptyset$.

The following concept is classic in Graph theory and we introduce it for completeness.

Bottom strongly connected component of a directed graph. In a directed graph, a bottom strongly connected component is a set of states where: there is a directed path from every state to every other state in the set, and all edges starting in the set lead to states within the set.

3.2 UPPER BOUND

The NP upper bound complexity is our main result and is established through the following sequence of results.

1. We establish a reduction from general POMDPs to blind MDPs (Lemma 1).
2. For blind MDPs, we establish the existence of Puiseux function policy witnesses (Lemma 2).
3. We establish the laminar structure of a graph of communicating classes in Puiseux Markov chains (Lemma 3).

4. We establish that the graph of communicating classes characterizes reachable states in Puiseux Markov chains (Lemma 4).
5. We establish properties of the graph of communicating classes that characterize limit-sure reachability in Puiseux Markov chains (Lemma 5).
6. We establish the existence of rank policy witnesses, a simple and polynomial-size policy (Lemma 6).
7. We provide a polynomial-time verifier for rank policy witnesses (Lemma 7).

The following result establishes a reduction from general POMDPs to blind MDPs.

Lemma 1. *For every POMDP $P = (\mathcal{S}, \mathcal{A}, \delta, \mathcal{Z}, o, s_0)$, there exists a blind MDP with $P' = (\mathcal{S}, \mathcal{A} \times \mathcal{Z}, \delta', s_0)$ with the same reachability value under memoryless policies.*

Proof sketch. The action (a, z) in the blind MDP corresponds to applying action a only to states whose observation is z. We define the transition δ' accordingly by introducing loops when an action (a, z) is applied and the underlying state s has a different observation, i.e., $z \neq o(s)$. A coupling on the underlying dynamics, that eliminates the introduced loops in the blind MDP, shows that the reachability values under memoryless policies coincide. □

Puiseux function policy. A (memoryless) Puiseux function policy σ is a function $\sigma \colon [0, \varepsilon_0) \to \Delta(\mathcal{A})$. Note that, for all $\varepsilon \in [0, \varepsilon_0)$, the policy $\sigma(\varepsilon)$ is a memoryless policy and, together with an initial state s, induces a Markov chain whose measure is denoted $\mathbb{P}_s^{\sigma(\varepsilon)}$. For example, for some $a \in \mathcal{A}$, we may have that $\sigma(\varepsilon)_a = 1/(2 - \varepsilon)$, which is a probability for $\varepsilon \in [0, 1]$.

The following result establishes the existence of Puiseux function policy witnesses for blind MDPs.

Lemma 2. *Consider a blind MDP $P = (\mathcal{S}, \mathcal{A}, \delta, s_0)$ and a target state $\top \in \mathcal{S}$. Then, P is limit-sure winning under memoryless policies if and only if the following decision problem for the first-order theory of the reals has a solution in the real-closed field of Puiseux functions.*

$\forall \lambda < 1\ \exists (\sigma_a)_{a \in \mathcal{A}}\ \exists (v_s)_{s \in \mathcal{S}}$ *such that*

- *Policy: for all $a \in \mathcal{A}$, we have that $\sigma_a \geq 0$, and $\sum_{a \in \mathcal{A}} \sigma_a = 1$.*
- *Fixpoint: for all $s \in \mathcal{S}$, we have that v satisfies*

$$v_s = \sum_{\tilde{s} \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sigma_a \delta(s, a)(\tilde{s})\, v_{\tilde{s}}\,.$$

- *Minimal solution: $\forall (u_s)_{s \in \mathcal{S}}$, if u satisfies the previous fixpoint equation, then, for all $s \in \mathcal{S}$, $v_s \leq u_s$.*
- *Value: $v_{s_0} \geq \lambda$.*

Proof sketch. We follow an approach similar to Bewley and Kohlberg [1976] where we characterize the limit-sure winning problem as a decision problem in the first-order theory of the reals: the POMDP is limit-sure winning if and only if this decision problem is true. By Tarski's principle [Basu et al., 2006, Theorem 2.80, page 70], the decision problem is true if and only if it has a witness in the field of Puiseux functions, which is a real-closed field. Therefore, limit-sure winning POMDPs have Puiseux functions policy witnesses. □

Graph of communicating classes. A memoryless Puiseux function policy σ on a blind MDP induces a Puiseux Markov chain, which defines communicating classes. The graph of communicating classes is a directed graph with one vertex per communicating class and an edge between two classes if the support of the exit distribution of one class contains a state in the other. Formally, consider $G = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{\mathcal{B} \subseteq \mathcal{S} : \mathcal{B} \text{ is a communicating class }\}$ and $(\mathcal{B}, \tilde{\mathcal{B}}) \in \mathcal{E}$ if and only if $\operatorname{supp}(\delta(\mathcal{B}, \cdot)) \cap \tilde{\mathcal{B}} \neq \emptyset$. The graph of communicating classes for Example 1, induced by the Puiseux policy σ, where $\sigma_\varepsilon(w) = 1 - \varepsilon$ and $\sigma_\varepsilon(c) = \varepsilon$, is illustrated in Figure 2.

The following result shows that communicating classes have a laminar structure.

Lemma 3. *Consider a Puiseux Markov chain with a set of states $\mathcal{S}$ and disjoint communicating classes $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k \subseteq \mathcal{S}$, with $k \geq 2$. We have that $\mathcal{B} := \sqcup_{i \in [k]} \mathcal{B}_i$ is a communicating class if and only if $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$ is a bottom strongly connected component in the graph with vertices $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\} \sqcup \{\bot\}$ and edges*

$$\{(\mathcal{B}_i, \mathcal{B}_j) : \exists s \in \mathcal{B}_j,\ s \in \operatorname{supp}(\delta(\mathcal{B}_i, \cdot))\} \sqcup$$
$$\{(\mathcal{B}_i, \bot) : \exists s \in \mathcal{S} \setminus \mathcal{B},\ s \in \operatorname{supp}(\delta(\mathcal{B}_i, \cdot))\}\,.$$

Proof sketch. The graph of the statement considers edges based on the exit distribution of communicating classes. On the one hand, if $\mathcal{B}$ is a communicating class, then starting in $\mathcal{B}$ the dynamic reaches every other state in $\mathcal{B}$ before exiting it. In particular, the exit distribution connects communicating classes between each other without leading to states outside. On the other hand, if $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$ is a bottom strongly connected component, then exit distributions link the classes, ensuring mutual reachability. Therefore, starting in a state in $\mathcal{B}$ the dynamic reaches every other state in $\mathcal{B}$ before exiting it, so $\mathcal{B}$ is a communicating class. □

The laminar structure implies the following bound on the number of communicating classes.

Corollary 1. *Consider a Puiseux Markov chain with a set of states $\mathcal{S}$. There are at most $2|\mathcal{S}| - 1$ communicating classes.*

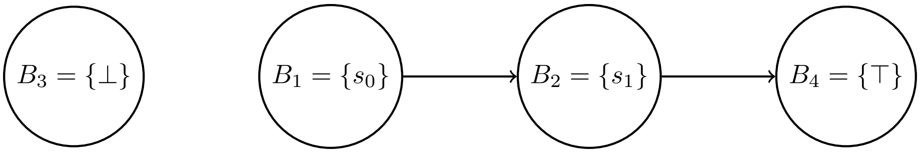

Figure 2: Graph of communicating classes induced by the Puiseux policy σ, where $\sigma_\varepsilon(w) = 1 - \varepsilon$ and $\sigma_\varepsilon(c) = \varepsilon$, for Example 1. Each node represents a communicating class in the induced Puiseux Markov chain. Directed edges denote non-zero exit probabilities (in the limit $\varepsilon \to 0$) between classes.

The following result characterizes reachable states in a Puiseux Markov chain.

Lemma 4. *Consider a Puiseux Markov chain with a set of states $\mathcal{S}$. The limit reachability is given by connectivity in the graph of communicating classes as follows. For all states $s, \tilde{s} \in \mathcal{S}$, we have that*

$$\lim_{\varepsilon \to 0^+} \mathbb{P}_s^\varepsilon(\mathit{Reach}(\tilde{s})) > 0$$

if and only if $\{s\}$ is connected to a communicating class $\mathcal{B} \ni \tilde{s}$ in the graph of communicating classes.

Proof. Note that the set of all reachable states in the limit from s, i.e.,

$$\left\{ \tilde{s} \in \mathcal{S} : \lim_{\varepsilon \to 0^+} \mathbb{P}_s^\varepsilon(\text{Reach}(\tilde{s})) > 0 \right\} ,$$

is characterized as the outcome of the following procedure, which is similar to the procedure used in the proof of Lemma 3. Start from $\{s\}$. First, closure by communicating class, if a state is included and this state is in some communicating class, then all states in the communicating class must be included. Second, closure by exit distribution, if a state is in the support of the exit distribution of a reachable communicating class, then it also must be included. Repeat the first and second closures until no more states are included to obtain the set of all reachable states in the limit from s. In particular, $\tilde{s}$ is reachable in the limit from s if and only if the communicating class $\{s\}$ is connected to $\mathcal{B} \ni \tilde{s}$ through, for example, a minimal sequence of additions in this process to include $\tilde{s}$ as a reachable state from s. □

The following result characterizes limit-sure reachability in Puiseux Markov chains.

Lemma 5. *A Puiseux Markov chain, with a set of states $\mathcal{S}$ and a reachability objective, is limit-sure winning starting from $s \in \mathcal{S}$ if and only if, for all communicating classes $\mathcal{B} \subseteq \mathcal{S}$, if $\{s\}$ is connected to $\mathcal{B}$ in the graph of communicating classes, then $\mathcal{B} = \{\top\}$ or the support of its exit distribution is not empty, i.e.,* $\text{supp}(\delta(\mathcal{B}, \cdot)) \neq \emptyset$.

Proof sketch. On the one hand, if P is limit-sure winning, then, by Lemma 4, $\{s\}$ is connected to $\{\top\}$ in the graph of communicating classes. By contradiction, if $\{s\}$ is connected to $\mathcal{B}$ and its exit distribution has empty support, then starting from s the dynamic has positive probability of reaching and staying forever in $\mathcal{B}$, which contradicts the limit-sure winning property. On the other hand, if s satisfies the assumptions, then we show that the dynamic eventually exits every subset of states containing s and not $\top$. Therefore, the dynamic reaches $\top$ with probability one in the limit, which proves that P is limit-sure winning. □

Rank policy witness. Given a blind MDP P, we say that a Puiseux policy σ is a *witness* for limit-sure winning if

$$\lim_{\varepsilon \to 0^+} \mathbb{P}_{s_0}^{\sigma(\varepsilon)}(\text{Reach}(\top)) = 1 \,.$$

A (memoryless) Puiseux policy $\sigma\colon [0, \varepsilon_0) \to \Delta(\mathcal{A})$ is a *rank* policy if, for all $a \in \mathcal{A}$, the function $\varepsilon \mapsto \sigma(\varepsilon)(a)$ is either constant to zero or an integer power of the identity up to normalization, i.e., there exists a function $f\colon \mathcal{A} \times [0, \varepsilon_0) \to [0, 1]$ such that, for all $a \in \mathcal{A}$, there exists $i \geq 0$ such that $f(a, \varepsilon) = \varepsilon^i$, and $\sigma(\varepsilon)(a) = f(a, \varepsilon) / \sum_{a \in \mathcal{A}} f(a, \varepsilon)$. In particular, for rank policies, we have that $\varepsilon_0 = \infty$.

The following result shows the existence of rank policy witnesses.

Lemma 6. *Consider a blind MDP $P = (\mathcal{S}, \mathcal{A}, \delta, s_0)$ and a target state $\top$. Then, P is limit-sure winning under memoryless policies if and only if there is a rank policy witness. Moreover, the description of the rank policy is of polynomial size.*

Proof sketch. By Lemma 2, P is limit-sure winning under memoryless policies if and only if there is a Puiseux policy witness. By Lemma 5, a Puiseux policy is a witness if and only if the respective graph of communicating classes has some properties. Note that the graph of communicating classes is defined only through the asymptotic behavior of the corresponding Puiseux Markov chain. Taking the communicating classes and the edges between them as a system of linear inequalities, we show the existence of a rank policy that induces the same graph of communicating classes and therefore is also a witness. Because ranks are the solution of a system of linear equations, they are of polynomial size. □

Lemma 6 establishes the existence of a polynomial-size witness for limit-sure reachability of a blind MDP. To prove the problem is in NP, the following result shows the existence of a polynomial-time verifier that decides whether a rank policy is a witness of limit-sure reachability for a blind MDP or not.

Lemma 7. *There exists a polynomial-time algorithm that, given a blind MDP and a rank policy, decides whether the rank policy is a witness of limit-sure reachability or not.*

Proof sketch. The algorithm has two main steps. First, given a rank policy, it constructs its graph of communicating classes. Second, it checks whether the only absorbing communicating class reachable from the initial state s_0 is $\{\top\}$ or not. The graph is constructed iteratively following the proof of Lemma 6. The main operations are adding edges and communicating classes. Each of them take polynomial time and, by Corollary 1, there are at most $(2|\mathcal{S}| - 1)$ communicating classes. For the second step, by the characterization in Lemma 5, we can decide limit-sure reachability running a depth-first search starting at $\{s_0\}$. Therefore, the algorithm runs in polynomial time. □

3.3 LOWER BOUND

An NP-hardness result was established for a similar problem by [Chatterjee et al., 2013, Lemma 1], namely, it was shown that the problem of determining whether a two-player game with partial-observation with reachability objective is limit-sure winning under memoryless policies is NP-hard. The reduction constructed a game that is a directed acyclic graph, and replacing the adversarial player with a uniform distribution over choices shows that the limit-sure winning problem under memoryless policies in POMDPs is also NP-hard.

Proposition 1. *For all constants $m \geq 0$, the problem of determining whether a POMDP P with reachability objective is limit-sure winning under memory m policies is NP-hard.*

We finish this section with the proof of Theorem 1.

Proof of Theorem 1. Proposition 1 establish the NP-hardness. Lemma 6 and Lemma 7 imply the existence of a rank policy of polynomial size and a polynomial-time verifier, which yields the NP upper bound. □

3.4 EXTENSIONS

In this section, we discuss several extensions of Theorem 1. The following result shows that Theorem 1 extends to constant memory policies.

Corollary 2. *The problem of determining whether a POMDP $P = (\mathcal{S}, \mathcal{A}, \delta, \mathcal{Z}, o, s_0)$ with reachability objective is limit-sure winning under constant memory policies is NP-complete.*

Proof sketch. The NP-hardness follows from Proposition 1. The NP upper bound is obtained as follows. For an amount of memory $m \geq 1$, guessing the update function σ_u, we can solve the memoryless problem on the product of the POMDP and the memory elements. Hence, the NP upper bound follows. □

Remark 1. *While Corollary 2 is stated for constant memory, the result holds for all memory bounds that are polynomial in the size of the POMDP, as this ensures that the witness is of polynomial size.*

While Corollary 2 presents the extension to small memory policies, we further extend Corollary 2 to other classic objectives, namely, parity or omega-regular objectives. Parity objectives are canonical forms to express all ω-regular properties [Thomas, 1997], e.g., all properties expressed in the linear-temporal logic (LTL) can be expressed as deterministic parity automata. In a parity condition, every state is labeled with a positive integer priority and the objective requires that the minimum priority visited infinitely often is even. For any fixed memory policy, we obtain a Markov chain, and the recurrent classes are reached with probability 1. A recurrent class satisfies the parity objective with probability 1 if the minimum priority is even, which we refer to as a good recurrent class, otherwise satisfies the objective with probability 0. Hence, the limit-sure problem for parity under memoryless strategies reduces to limit-surely reaching the good recurrent classes. Hence, the NP-completeness result of Theorem 1 and Corollary 2 also extend to parity objectives. However, we focus on reachability objectives as all conceptual aspects are clarified in this basic and most fundamental objective.

Corollary 3. *The problem of determining whether a POMDP P with parity objective is limit-sure winning under constant memory policies is NP-complete.*

Proof sketch. The NP-hardness follows from Proposition 1. The NP upper bound is obtained as follows. By guessing the support of rank policies, we can compute the recurrent classes, and the objective is to reach the good recurrent classes. □

Remark 2. *As mentioned before, while Corollary 3 is stated for constant memory, the result holds for all memory bounds that are polynomial in the POMDP size, as this ensures that the witness is of polynomial size.*

The following result shows that Theorem 1 extends to parametric Markov chains (pMCs) as presented by [Junges et al., 2018].

Corollary 4. *The problem of determining whether a parametric Markov chain with reachability objective is limit-sure winning under constant memory policies is NP-complete.*

Proof sketch. By [Junges et al., 2018, Corollary 1], the quantitative problem for POMDPs and pMCs for reach-avoid objectives are equivalent. In particular, the NP upper bound for limit-sure winning for reachability objectives translates immediately. □

Corollary 4 complements the qualitative objectives investigated by [Junges et al., 2021], which include almost-sure but not limit-sure reachability.

4 CONCLUSION AND FUTURE WORK

In this work, we presented the first solution for limit-sure winning with small memory policies for POMDPs. While the present work establishes the theoretical foundations, interesting directions of future work include the development of efficient encodings in NP-complete problems that have well developed solvers. This includes SAT and Mixed Linear Program (MLP).

Along evaluating combinations of reductions and solvers in standard benchmark instances, an important task is to identify classes of POMDPs on which a solver works particularly well, possibly including efficient heuristics for scalability and practical applications. Besides classic reductions, incremental encodings should be investigated. In other words, generating the clauses for SAT and the restrictions for MLP incrementally as opposed to generating all of them at the same time. Incremental encodings have been developed for almost-sure reachability, see for example [Chatterjee et al., 2016], and they take advantage of incremental solvers obtaining meaningful practical improvements.

Acknowledgements

This research was partially supported by Austrian Science Fund (FWF) 10.55776/COE12, the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-21-CE40-0020 (CONVERGENCE project), and the ERC CoG 863818 (ForM-SMArt) grant.

References

Robert J . Aumann. Mixed and Behavior Strategies in Infinite Extensive Games. In *Advances in Game Theory. (AM-52)*, pages 627–650. Princeton University Press, 1964. doi: 10.1515/9781400882014-029.

Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. MIT Press, 2008.

Christel Baier, Marcus Grösser, and Nathalie Bertrand. Probabilistic ω-Automata. *Journal of the ACM*, 59(1):1–52, February 2012. doi: 10.1145/2108242.2108243.

S. Baruah, G. Koren, D. Mao, B. Mishra, A. Raghunathan, L. Rosier, D. Shasha, and F. Wang. On the Competitiveness of On-Line Real-Time Task Scheduling. *Real-Time Systems*, 4(2):125–144, 1992. doi: 10.1007/BF00365406.

Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry*, volume 10 of *Algorithms and Computation in Mathematics*. Springer, 2006. doi: 10.1007/3-540-33099-2.

Richard Bellman. A Markovian Decision Process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957.

Dimitri P. Bertsekas. *Dynamic Programming and Stochastic Control*. Number v. 125 in Mathematics in Science and Engineering. Academic Press, New York, 1976.

Truman Bewley and Elon Kohlberg. The Asymptotic Theory of Stochastic Games. *Mathematics of Operations Research*, 1(3):197–208, 1976. doi: 10.1287/moor.1.3.197.

Patrick Billingsley. *Probability and Measure*. Wiley, 2012.

Pavol Černý, Krishnendu Chatterjee, Thomas A. Henzinger, Arjun Radhakrishna, and Rohit Singh. Quantitative Synthesis for Concurrent Programs. In Ganesh Gopalakrishnan and Shaz Qadeer, editors, *Computer Aided Verification*, volume 6806, pages 243–259. Springer, 2011. doi: 10.1007/978-3-642-22110-1_20.

Krishnendu Chatterjee and Thomas A. Henzinger. Probabilistic Automata on Infinite Words: Decidability and Undecidability Results. In *Automated Technology for Verification and Analysis*, volume 6252, pages 1–16. Springer, 2010. doi: 10.1007/978-3-642-15643-4_1.

Krishnendu Chatterjee, Laurent Doyen, and Thomas A. Henzinger. Qualitative Analysis of Partially-Observable Markov Decision Processes. In *Mathematical Foundations of Computer Science (MFCS)*, volume 6281, pages 258–269. Springer, 2010. doi: 10.1007/978-3-642-15155-2_24.

Krishnendu Chatterjee, Alexander Kößler, and Ulrich Schmid. Automated Analysis of Real-Time Scheduling Using Graph Games. In *Proceedings of the 16th International Conference on Hybrid Systems: Computation and Control*, pages 163–172, 2013. doi: 10.1145/2461328.2461356.

Krishnendu Chatterjee, Martin Chmelik, Raghav Gupta, and Ayush Kanodia. Qualitative Analysis of POMDPs with Temporal Logic Specifications for Robotics Applications. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 325–330, 2015. doi: 10.1109/ICRA.2015.7139019.

Krishnendu Chatterjee, Martin Chmelík, and Jessica Davies. A Symbolic SAT-Based Algorithm for Almost-Sure Reachability with Small Strategies in POMDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30 (1):3225–3232, 2016. doi: 10.1609/aaai.v30i1.10422.

Karel Culik and Jarkko Kari. Digital Images and Formal Languages. In Grzegorz Rozenberg and Arto Salomaa, editors, *Handbook of Formal Languages*, pages 599–616. Springer, 1997. doi: 10.1007/978-3-642-59126-6_10.

Edsger W. Dijkstra. *A Discipline of Programming*. Prentice Hall series in automatic computation. Prentice Hall, Englewood Cliffs, NJ, 1976.

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1 edition, 1998. doi: 10.1017/CBO9780511790492.

Richard Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, fifth edition edition, 2019. doi: 10.1017/9781108591034.

Eugene A. Feinberg. On Measurability and Representation of Strategic Measures in Markov Decision Processes. In *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pages 29–43. Institute of Mathematical Statistics, 1996. doi: 10.1214/lnms/1215453563.

Jerzy Filar and Koos Vrieze. *Competitive Markov Decision Processes*. Springer, New York, NY, 1997. doi: 10.1007/978-1-4612-4054-9.

Harold N. Gabow, Zvi Galil, Thomas Spencer, and Robert E. Tarjan. Efficient Algorithms for Finding Minimum Spanning Trees in Undirected and Directed Graphs. *Combinatorica*, 6(2):109–122, 1986. doi: 10.1007/BF02579168.

Hugo Gimbert and Youssouf Oualhadj. Probabilistic Automata on Finite Words: Decidable and Undecidable Problems. In *Automata, Languages and Programming*, volume 6199, pages 527–538. Springer, 2010. doi: 10.1007/978-3-642-14162-1_44.

Ronald A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, 1960.

Sebastian Junges, Nils Jansen, Ralf Wimmer, Tim Quatmann, Leonore Winterer, Joost-P Katoen, and Bernd Becker. Finite-State Controllers of POMDPs Using Parameter Synthesis. In *Uncertainty in Artificial Intelligence*, pages 519–529, 2018.

Sebastian Junges, Joost-Pieter Katoen, Guillermo A. Pérez, and Tobias Winkler. The Complexity of Reachability in Parametric Markov Decision Processes. *Journal of Computer and System Sciences*, 119:183–210, 2021. doi: 10.1016/j.jcss.2021.02.006.

Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4(1):237–285, 1996.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and Acting in Partially Observable Stochastic Domains. *Artificial Intelligence*, 101(1-2): 99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Richard M. Karp. Reducibility among Combinatorial Problems. In *Complexity of Computer Computations: Proceedings of a Symposium on the Complexity of Computer Computations*, pages 85–103. Springer, 1972. doi: 10.1007/978-1-4684-2001-2_9.

H. Kress-Gazit, G.E. Fainekos, and G.J. Pappas. Temporal-Logic-Based Reactive Mission and Motion Planning. *IEEE Transactions on Robotics*, 25(6):1370–1381, December 2009. doi: 10.1109/TRO.2009.2030225.

Omid Madani, Steve Hanks, and Anne Condon. On the Undecidability of Probabilistic Planning and Related Stochastic Optimization Problems. *Artificial Intelligence*, 147(1):5–34, July 2003. doi: 10.1016/S0004-3702(02)00378-8.

Mehryar Mohri. Finite-State Transducers in Language and Speech Processing. *Computational Linguistics*, 23(2): 269–311, June 1997.

Christos H. Papadimitriou and John N. Tsitsiklis. The Complexity of Markov Decision Processes. *Mathematics of Operations Research*, 12(3):441–450, 1987. doi: 10.1287/moor.12.3.441.

Azaria Paz and Werner Rheinboldt. *Introduction to Probabilistic Automata*. Computer Science and Applied Mathematics. Academic Press, 1971.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 2014.

Eilon Solan. Perturbations of Markov Chains with Applications to Stochastic Games. In *Stochastic Games and Applications*, pages 265–280, 2003. doi: 10.1007/978-94-010-0189-2_17.

Wolfgang Thomas. Languages, Automata, and Logic. In *Handbook of Formal Languages*, pages 389–455. Springer, 1997. doi: 10.1007/978-3-642-59126-6_7.

Xavier Venel and Bruno Ziliotto. Strong Uniform Value in Gambling Houses and Partially Observable Markov Decision Processes. *SIAM Journal on Control and Optimization*, 54(4):1983–2008, 2016. doi: 10.1137/15M1043340.

Limit-sure Reachability for Small Memory Policies in POMDPs is NP-complete (Supplementary Material)

Ali Asadi[1] **Krishnendu Chatterjee**[1] **Raimundo Saona**[1] **Ali Shafiee**[1]

[1]Insitute of Science and Technology Austria, Klosterneuburg, Austria

A PROOFS OF SECTION 3.2

In this section, we provide the detailed proofs of results in Section 3.2.

Lemma (Restated, Lemma 1)**.** *For every POMDP $P = (\mathcal{S}, \mathcal{A}, \delta, \mathcal{Z}, o, s_0)$, there exists a blind MDP with $P' = (\mathcal{S}, \mathcal{A} \times \mathcal{Z}, \delta', s_0)$ with the same reachability value under memoryless policies.*

Proof. Consider an arbitrary POMDP $P = (\mathcal{S}, \mathcal{A}, \delta, \mathcal{Z}, o, s_0)$. Define a blind MDP $P' = (\mathcal{S}, \mathcal{A}', \delta', \mathcal{Z}', o', s_0)$ where

- $\mathcal{A}' := \mathcal{A} \times \mathcal{Z}$;
- $\delta' \colon \mathcal{S} \times \mathcal{A}' \to \Delta(\mathcal{S})$ is given by

$$\delta'(s, (a, z)) := \begin{cases} \delta(s, a) & o(s) = z \\ \mathbb{1}[s] & o(s) \neq z \end{cases}$$

- $\mathcal{Z}' := \{\#\}$ a unique observation;
- $o' \equiv \#$ a uninformative observation function.

We show that the value of this blind MDP P' is the same as the original POMDP P.

Consider an arbitrary memoryless policy $\sigma \colon \mathcal{Z} \to \Delta(\mathcal{A})$ in the POMDP P. Note that $\sigma \in \Delta(\mathcal{A})^{\mathcal{Z}}$ is a collection of distributions. Define the memoryless policy $\sigma' \colon \mathcal{Z}' \to \Delta(\mathcal{A}')$ in P', which we identify with an element of $\Delta(\mathcal{A}')$, by a uniform choice over the distributions in σ, i.e.,

$$\sigma'((a, z)) := \frac{1}{|Z|} \sigma(z)(a) \, .$$

In other words, the policy σ' chooses an observation z uniformly at random and then an action according to the distribution $\sigma(z)$.

The coupling between the blind MDP and the POMDP consists in projecting the dynamic of the blind MDP to those times where the tuple of action and observation (a, z) is such that z is the observation of the current state. Formally, define a sequence of random times $(\tau_t)_{t \geq 0}$ defined inductively by $\tau_0 := \inf\{t \geq 0 : \exists a \in \mathcal{A} \quad A'_t = (a, o(s_0))\}$ and, for $t \geq 1$,

$$\tau_t := \inf\{t > \tau_{t-1} : \exists a \in \mathcal{A} \quad A'_t = (a, o(S_{t-1}))\} \, .$$

In other words, τ_t is the t-th time that, in the dynamic of the blind MDP, the second coordinate of the action chosen by σ' coincides with the observation of the current state in the original POMDP. These times are almost surely finite since σ' chooses an observation uniformly at random at each step. Notice that, after coupling the transitions of σ and σ' in the obvious way, $(S_{\tau_t})_{t \geq 0}$ under σ' and $(S_t)_{t \geq 0}$ under σ follow the same dynamic. In particular, the probability of reaching the target is equal in the blind MDP and the POMDP. Therefore, the reachability value of the blind MDP is at least as large as the reachability value of the POMDP.

Consider an arbitrary memoryless policy $\sigma' \colon \mathcal{Z}' \to \Delta(\mathcal{A}')$ in the blind MDP P', or equivalently an element of $\Delta(\mathcal{A}')$. Define the memoryless policy $\sigma \colon \mathcal{Z} \to \Delta(\mathcal{A})$ in P by

$$\sigma(z)(a) := \frac{\sigma'((a, z))}{\sum_{\tilde{a} \in \mathcal{A}} \sigma'((\tilde{a}, z))} \,.$$

In other words, for each observation, we consider the conditional distribution of σ' on the actions that have that observation as a second coordinate.

Just as before, the same coupling shows $(S_{\tau_t})_{t \geq 0}$ under σ' and $(S_t)_{t \geq 0}$ under σ follow the same dynamic. Therefore, the reachability value of the POMDP is at least as large as the reachability value of the blind MDP. We conclude that both POMDPs have the same value. □

Lemma (Restated, Lemma 2). *Consider a blind MDP $P = (\mathcal{S}, \mathcal{A}, \delta, s_0)$ and a target state $\top \in \mathcal{S}$. Then, P is limit-sure winning under memoryless policies if and only if the following decision problem for the first-order theory of the reals has a solution in the real-closed field of Puiseux functions.*

$\forall \lambda < 1 \; \exists (\sigma_a)_{a \in \mathcal{A}} \; \exists (v_s)_{s \in \mathcal{S}}$ *such that*

- *Policy: for all $a \in \mathcal{A}$, we have that $\sigma_a \geq 0$, and $\sum_{a \in \mathcal{A}} \sigma_a = 1$.*
- *Fixpoint: for all $s \in \mathcal{S}$, we have that v satisfies*

$$v_s = \sum_{\tilde{s} \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sigma_a \delta(s, a)(\tilde{s}) \, v_{\tilde{s}} \,.$$

- *Minimal solution: $\forall (u_s)_{s \in \mathcal{S}}$, if u satisfies the previous fixpoint equation, then, for all $s \in \mathcal{S}$, $v_s \leq u_s$.*
- *Value: $v_{s_0} \geq \lambda$.*

Proof. Consider a blind MDP P. Recall that P is limit-sure winning under memoryless policies if and only if

$$\sup_{\sigma \in \Sigma_0} \mathbb{P}^{\sigma}_{s_0}(\text{Reach}(\top)) = 1 \,.$$

In other words, if and only if, for all $\lambda < 1$ there exists a policy $\sigma_\lambda \in \Sigma_0$ such that $\mathbb{P}^{\sigma}_{s_0}(\text{Reach}(\top)) \geq \lambda$. Recall that, since P is a blind MDP, the set of memoryless policies Σ_0 is equivalent to $\Delta(\mathcal{A})$, so an element $\sigma \in \Sigma_0$ is fully determined by the probability it assigns to each action.

By Theorem 4, a policy $\sigma_\lambda \in \Sigma_0$ is such that $\mathbb{P}^{\sigma_\lambda}_{s_0}(\text{Reach}(\top)) \geq \lambda$ if and only if the corresponding Markov chain has a value vector such that $v_{s_0} = v(s_0) \geq \lambda$. So far, we conclude that P is limit-sure winning under memoryless policies if and only if the stated decision problem for the first-order theory of the reals has a solution in $\mathbb{R}$. By Theorem 3, since the Puiseux functions is a real-closed field by Theorem 2 and it contains $\mathbb{R}$, we conclude the proof. □

Lemma (Restated, Lemma 3). *Consider a Puiseux Markov chain with a set of states $\mathcal{S}$ and disjoint communicating classes $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k \subseteq \mathcal{S}$, with $k \geq 2$. We have that $\mathcal{B} := \sqcup_{i \in [k]} \mathcal{B}_i$ is a communicating class if and only if $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$ is a bottom strongly connected component in the graph with vertices $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\} \sqcup \{\bot\}$ and edges*

$$\begin{gathered}\{(\mathcal{B}_i, \mathcal{B}_j) : \exists s \in \mathcal{B}_j, \; s \in \operatorname{supp}(\delta(\mathcal{B}_i, \cdot))\} \sqcup \\ \{(\mathcal{B}_i, \bot) : \exists s \in \mathcal{S} \setminus \mathcal{B}, \; s \in \operatorname{supp}(\delta(\mathcal{B}_i, \cdot))\} \,.\end{gathered}$$

Proof. Assume that $\mathcal{B}$ is a communicating class. We show that $\mathcal{B}$ is a bottom strongly connected component in the graph of the statement. Consider $i \in [k]$ arbitrary. We show that all edges of $\mathcal{B}_i$ lead to communicating classes in $\mathcal{B}$, i.e., $\operatorname{supp}(\delta(\mathcal{B}_i, \cdot)) \subseteq \mathcal{B}$. By contradiction, assume that $\tilde{s} \in \operatorname{supp}(\delta(\mathcal{B}_i, \cdot)) \cap \mathcal{S} \setminus \mathcal{B}$, equivalently, the graph of the statement contains an edge $(\mathcal{B}_i, \bot)$. Because $\mathcal{B}$ is a communicating class and $k \geq 2$, there exists $j \neq i$ and $\mathcal{B}_j$ such that $(\mathcal{B}_i, \mathcal{B}_j)$ is an edge in the graph of the statement. Consider states $s \in \mathcal{B}_i$ and $\tilde{\tilde{s}} \in \mathcal{B}_j$ where $\tilde{\tilde{s}}$ is such that $\delta(\mathcal{B}_i, \tilde{\tilde{s}}) > 0$. On the one hand, because $\mathcal{B}$ is a communicating class, $\tilde{s}$ is reached starting from s before exiting $\mathcal{B}$, i.e.,

$$\lim_{\varepsilon \to 0^+} \mathbb{P}^{\varepsilon}_{s}(\text{exit}(B) < \text{reach}(\tilde{\tilde{s}})) = 0 \,.$$

On the other hand, by definition of exit distribution, there is a positive limit probability to exit $\mathcal{B}_i$ through $\tilde{s}$ which is outside of $\mathcal{B}$, i.e.,

$$\lim_{\varepsilon \to 0^+} \mathbb{P}^{\varepsilon}_{s}(\text{Exit}(B, \tilde{s})) \geq \delta(\mathcal{B}_i, \tilde{s}) > 0\,.$$

This is a contradiction. Therefore, $\operatorname{supp}(\delta(\mathcal{B}_i, \cdot)) \subseteq \mathcal{B}$. We are left with showing that $\mathcal{B}$ is strongly connected in the graph of the statement.

Consider $i \neq j \in [k]$ arbitrary. We show that $\mathcal{B}_i$ is connected to $\mathcal{B}_j$ in the graph of the statement. Consider $s \in \mathcal{B}_i$ and $\tilde{s} \in \mathcal{B}_j$. On the one hand, because $\mathcal{B}$ is a communicating class, $\tilde{s}$ is reached starting from s before exiting $\mathcal{B}$. On the other hand, the set of all reachable states from s before having left $\mathcal{B}$ is characterized by the following procedure. Start including s. First, closure by communicating class, if $\tilde{\tilde{s}}$ is reachable from s and $\tilde{\tilde{s}} \in \mathcal{B}_\ell$ with $\ell \in [k]$, then all states in $\mathcal{B}_\ell$ are included. Second, closure by exit distribution, if a state is in the support of the exit distribution of the reachable communicating classes, then it also is included. Repeat these closures until no more states are included to obtain the set of all reachable states from s before having left $\mathcal{B}$. In particular, $\tilde{s}$ must be included in one of these two closures, and $\mathcal{B}_i$ is connected to $\mathcal{B}_j$ through, for example, a minimal sequence of additions in this process to include $\tilde{s}$ as a reachable state from s before having left $\mathcal{B}$. We conclude that $\mathcal{B}$ is a bottom strongly connected component in the graph of the statement.

Assume that $\mathcal{B}$ is a bottom strongly connected component in the graph of the statement. We show that $\mathcal{B}$ is a communicating class. Consider $s, \tilde{s} \in \mathcal{B}$. We show that the limit probability of starting at s and reaching $\tilde{s}$ before leaving $\mathcal{B}$ is 1. Consider $i, j \in [k]$ such that $s \in \mathcal{B}_i$ and $\tilde{s} \in \mathcal{B}_j$. By definition of communicating classes, transitions within a communicating class occur before exiting it, so it is sufficient to consider only the transitions exiting communicating classes. Because $\mathcal{B}$ is a bottom strongly connected component in the graph of the statement, the exit distribution of all communicating classes leads to states in $\mathcal{B}$. In particular, the transitions between the communicating classes are taken infinitely more often than those exiting $\mathcal{B}$. Therefore, it is enough to show that starting from $\mathcal{B}_i$ the probability of having visited $\mathcal{B}_j$ after k transitions between communicating classes is strictly positive. Denote the smallest positive exit probability $\nu \coloneqq \min\{\delta(\mathcal{B}_\ell, s) : \ell \in [k], s \in \mathcal{S}, \delta(\mathcal{B}_\ell, s) > 0\} > 0$. Because $\mathcal{B}$ is strongly connected in the graph of the statement, there is a directed path between $\mathcal{B}_i$ and $\mathcal{B}_j$. Then, starting from $\mathcal{B}_i$ the probability of having visited $\mathcal{B}_j$ after k transitions between communicating classes is at least $\nu^k > 0$. We conclude that $\mathcal{B}$ is a communicating class. □

Lemma (Restated, Lemma 5)**.** *A Puiseux Markov chain, with a set of states $\mathcal{S}$ and a reachability objective, is limit-sure winning starting from $s \in \mathcal{S}$ if and only if, for all communicating classes $\mathcal{B} \subseteq \mathcal{S}$, if $\{s\}$ is connected to $\mathcal{B}$ in the graph of communicating classes, then $\mathcal{B} = \{\top\}$ or the support of its exit distribution is not empty, i.e.,* $\operatorname{supp}(\delta(\mathcal{B}, \cdot)) \neq \emptyset$*.*

Proof. Consider a Puiseux Markov chain P and a state $s \in \mathcal{S}$. Assume that P is limit-sure winning starting from s. On the one hand,

$$\lim_{\varepsilon \to 0^+} \mathbb{P}^{\sigma(\varepsilon)}_{s}(\text{Reach}(\top)) = 1 > 0\,.$$

In particular, by Lemma 4, $\{s\}$ is connected to a communicating class $\mathcal{B} \ni \top$. Note that, by definition, $\mathcal{B} = \{\top\}$ is an absorbing communicating class. By Lemma 3, the only communicating class containing $\top$ is $\{\top\}$. We conclude that $\{s\}$ is connected to $\{\top\}$ in the graph of communicating classes. On the other hand, consider a communicating class $\mathcal{B}$ such that $\{s\}$ is connected to $\mathcal{B}$ in the graph of communicating classes. Consider $\tilde{s} \in \mathcal{B}$. By Lemma 4,

$$\lim_{\varepsilon \to 0^+} \mathbb{P}^{\varepsilon}_{s}(\text{Reach}(\tilde{s})) > 0\,.$$

By contradiction, if $\mathcal{B}$ is absorbing, then

$$\lim_{\varepsilon \to 0^+} \mathbb{P}^{\sigma(\varepsilon)}_{s}(\text{Reach}(\top)) \leq 1 - \lim_{\varepsilon \to 0^+} \mathbb{P}^{\varepsilon}_{s}(\text{Reach}(\tilde{s})) < 1\,,$$

which is a contradiction. We conclude that the support of the exit distribution of $\mathcal{B}$ is not empty.

Assume that, for all communicating classes $\mathcal{B} \subseteq \mathcal{S}$, if $\{s\}$ is connected to $\mathcal{B}$ in the graph of communicating classes, then $\mathcal{B} = \{\top\}$ or the support of its exit distribution is not empty, i.e., $\operatorname{supp}(\delta(\mathcal{B}, \cdot)) \neq \emptyset$. We show that P is limit-sure winning starting from s. Note that, starting from s, the dynamic either reaches $\top$ or stays in a subset of $\mathcal{S} \setminus \{\top\}$ forever. For a subset $\mathcal{C} \subseteq \mathcal{S}$, denote the time from which the dynamic never leaves $\mathcal{C}$ again by

$$\text{stay}(\mathcal{C}) \coloneqq \min\{n \geq 0 : \forall \tilde{n} \geq n \quad s_{\tilde{n}} \in \mathcal{C}\}\,.$$

We show that, for all $\mathcal{C} \subseteq \mathcal{S}$ such that $s \in \mathcal{C}$ and $\top \notin \mathcal{C}$, the probability of staying in $\mathcal{C}$ forever is zero, i.e.,

$$\lim_{\varepsilon \to 0^+} \mathbb{P}^{\sigma(\varepsilon)}_{s}(\text{stay}(\mathcal{C}) < \infty) = 0\,.$$

Fix an arbitrary $\mathcal{C} \subseteq \mathcal{S}$ such that $\top \notin \mathcal{C}$ and

$$\lim_{\varepsilon \to 0^+} \mathbb{P}_s^{\sigma(\varepsilon)}(\text{Reach}(\mathcal{C})) > 0 \,.$$

Take $\tilde{s} \in \mathcal{C}$ such that $\lim_{\varepsilon \to 0^+} \mathbb{P}_s^{\sigma(\varepsilon)}(\text{Reach}(\tilde{s})) > 0$. By Lemma 4, there are communicating classes $\mathcal{B}, \tilde{\mathcal{B}}$ such that $s \in \mathcal{B}$, $\tilde{s} \in \tilde{\mathcal{B}}$, and $\mathcal{B}$ is connected to $\tilde{\mathcal{B}}$ in the graph of communicating classes. If $\tilde{\mathcal{B}} \not\subseteq \mathcal{C}$, then, by definition of communicating classes, the dynamic exits $\mathcal{C}$ in finite time and derive the result. By contradiction, assume that all communicating classes reachable from $\mathcal{B}$ are contained in $\mathcal{C}$. Consider the graph of communicating classes restricted to the classes included in $\mathcal{C}$. Because this is a directed subgraph, it has a bottom strongly connected component reachable from $\mathcal{B}$. Consider $\tilde{\mathcal{B}}$ the union of all states in this bottom strongly connected component. By Lemma 3, $\tilde{\mathcal{B}}$ is a communicating class. By construction, $\{s_0\}$ is connected to $\tilde{\mathcal{B}}$. Because $\tilde{\mathcal{B}} \subseteq \mathcal{C}$, we have that $\top \notin \tilde{\mathcal{B}}$. Therefore, by assumption, the support of its exit distribution is not empty, i.e., $\text{supp}(\delta(\mathcal{B}, \cdot)) \neq \emptyset$. But this is a contradiction with being a bottom strongly connected component. We conclude that, for all $\mathcal{C} \subseteq \mathcal{S} \setminus \{\top\}$, if $\lim_{\varepsilon \to 0^+} \mathbb{P}_s^{\sigma(\varepsilon)}(\text{Reach}(\mathcal{C})) > 0$, then $\lim_{\varepsilon \to 0^+} \mathbb{P}_s^{\sigma(\varepsilon)}(\text{stay}(\mathcal{C}) < \infty) = 0$. Therefore, $\lim_{\varepsilon \to 0^+} \mathbb{P}_s^{\sigma(\varepsilon)}(\text{Reach}(\top)) = 1$ and the Puiseux Markov chain is limit-sure winning. □

Lemma (Restated, Lemma 6)**.** *Consider a blind MDP $P = (\mathcal{S}, \mathcal{A}, \delta, s_0)$ and a target state $\top$. Then, P is limit-sure winning under memoryless policies if and only if there is a rank policy witness. Moreover, the description of the rank policy is of polynomial size.*

Proof. Consider a blind MDP $P = (\mathcal{S}, \mathcal{A}, \delta, s_0)$ and a target state $\top$. By Lemma 2, P is limit-sure winning under memoryless policies if and only if there is a memoryless Puiseux function policy witness. In turn, by Lemma 5, a policy is a witness if and only if its graph of communicating classes satisfies some properties. We show that, if there is a Puiseux function policy whose graph satisfies these properties, then there is another polynomial-size rank policy that induces the same graph.

Fix a Puiseux policy $\sigma \colon [0, \varepsilon_0) \to \Delta(\mathcal{A})$ and consider the corresponding Puiseux Markov chain. We claim that the graph of communicating classes is fully determined by the support of the exit distribution of communicating classes. Indeed, this graph can be constructed as follows.

- Initialization. Start by considering a communicating class for each singleton state.
- Adding edges. Given the currently considered communicating classes, add all edges given by the support of their exit distribution.
- Adding new communicating classes. Given the currently considered communicating classes, consider those that are not contained in another communicating class. With the support of their exit distribution, by Lemma 3, we find larger communicating classes if there are any.

By repeating the last two items until no other communicating class is found, we obtain the full graph of communicating classes. Therefore, this graph is fully determined by the support of the exit distribution of communicating classes. By Theorem 6, the exit distribution of a communicating class is characterized in terms of exit graphs. Indeed, a state $\tilde{s}$ is in the support of the exit distribution of a communicating class $\mathcal{B}$ if and only if there exists a state $s \in \mathcal{B}$ and an exit graph g in which s can reach $\tilde{s}$, i.e., $g \in G_{\mathcal{B}}(s \to \tilde{s})$, such that the weight of g is equal to the minimal weight of all exit graphs of $\mathcal{B}$, i.e., for all $\tilde{g} \in G_{\mathcal{B}}$ we have that $w(g) \leq w(\tilde{g})$. We use this characterization to deduce the existence of a rank policy $\tilde{\sigma}$ that induces the same graph as the policy σ.

Consider a parameterized function $f \colon \mathcal{A} \times [0, \varepsilon_0) \to [0, 1]$, of parameters $(i(a))_{a \in \mathcal{A}}$, given by $f(a, \varepsilon) = \varepsilon^{i(a)}$. This function induces a policy $\tilde{\sigma}(\varepsilon)(a) = f(a, \varepsilon) / \sum_{a \in \mathcal{A}} f(a, \varepsilon)$, which in turn defines a Puiseux Markov chain with transitions

$$\begin{aligned} \delta^{\varepsilon}(s, \tilde{s}) &= \sum_{a \in \mathcal{A}} \tilde{\sigma}(\varepsilon)(a) \delta(s, a)(\tilde{s}) \\ &= \frac{1}{\sum_{a \in \mathcal{A}} f(a, \varepsilon)} \sum_{a \in \mathcal{A}} \varepsilon^{i(a)} \delta(s, a)(\tilde{s}) \,. \end{aligned}$$

We show that there exist parameters $(i(a))_{a \in \mathcal{A}}$ such that the corresponding graph of communicating classes of $\tilde{\sigma}$ coincides with the one given by σ. Indeed, because the definition of exit distribution depends only on the weight of exit graphs, which is an asymptotic notion, we have that the weights of the policy σ induce a strategy with the same graph of communicating classes, i.e., defining

$$i(a) := \inf \left\{ r \geq 0 : \lim_{\varepsilon \to 0^+} \frac{\sum_{a \in \mathcal{A}} \sigma(\varepsilon)(a)\delta(s,a)(\tilde{s})}{\varepsilon^r} \right\}$$

we have that $\tilde{\sigma}$ induces the same graph of communicating classes as σ. We show that the parameters $(i(a))_{a \in \mathcal{A}}$ can be chosen to be integers of polynomial size.

By the previous arguments, for simplicity and without loss of generality consider that the Puiseux policy σ is of the form $\sigma(\varepsilon)(a) = f(a,\varepsilon) / \sum_{a \in \mathcal{A}} f(a,\varepsilon)$, where $f \colon \mathcal{A} \times [0,\varepsilon_0) \to [0,1]$ is such that $f(a,\varepsilon) = \varepsilon^{i(a)}$. We construct a system of linear equations that is solved by $(i(a))_{a \in \mathcal{A}}$ and characterizes the induced graph of communicating classes. First, ranking of actions. Consider (strict) inequalities that characterize the order of $(i(a))_{a \in \mathcal{A}}$, i.e., (strict) inequalities of the form

$$i(a) < i(\tilde{a}) \qquad \text{or} \qquad i(a) = i(\tilde{a}) \,.$$

Second, support of exit distributions. Consider states $s, \tilde{s} \in \mathcal{S}$ and define the set actions that lead to the transition from s to $\tilde{s}$ and are minimal in the ranking, i.e.,

$$\begin{aligned} \mathcal{I}(s \to \tilde{s}) := \{ a \in \mathcal{A} : \delta(s,a)(\tilde{s}) > 0,\, &\forall \tilde{a} \in \mathcal{A}, \\ &\delta(s,a)(\tilde{s}) > 0 \Rightarrow i(a) \leq i(\tilde{a}) \} \,. \end{aligned}$$

Also, consider some selection and define $i(s \to \tilde{s}) := i(a)$, for some $a \in \mathcal{I}(s \to \tilde{s})$. Recalling that a state $\tilde{s}$ is in the support of the exit distribution of a communicating class $\mathcal{B}$ if and only if there exists a state $s \in \mathcal{B}$ and an exit graph g containing the edge $(s, \tilde{s})$, i.e., $g \in G_{\mathcal{B}}(s \to \tilde{s})$, such that the weight of g is equal to the minimal weight of all exit graphs of $\mathcal{B}$, i.e., for all $\tilde{g} \in G_{\mathcal{B}}$ we have that $w(g) \leq w(\tilde{g})$. We write these restrictions as (strict) inequalities over $(i(a))_{a \in \mathcal{A}}$ by noticing that

$$w(g) = \sum_{(s,\tilde{s}) \in g} i(s \to \tilde{s}) \,.$$

Because there are finitely many communicating classes, and each of them has finitely many exit graphs, the graph of communicating classes induced by the policy $\tilde{\sigma}$ is fully determined by a finite system of, possibly strict, inequalities over the variables $(i(a))_{a \in \mathcal{A}}$. These two sets of (strict) inequalities, along with positivity constraints, fully characterize the induced graph of communicating classes by $(i(a))_{a \in \mathcal{A}}$ in the following sense. Every solution of these inequalities $(i^*(a))_{a \in \mathcal{A}}$ define a function $f^* \colon \mathcal{A} \times [0, \varepsilon_0) \to [0,1]$, which defines a policy σ^*. By the iterative construction of the graph of communicating class, the policy σ^* induces the same graph as σ. We show that these inequalities have an integer solution $(i^*(a))_{a \in \mathcal{A}}$ of polynomial size.

For a fixed order over $(i(a))_{a \in \mathcal{A}}$ and a selection defining $(i(s \to \tilde{s}))_{s, \tilde{s} \in \mathcal{S}}$, the inequalities considered are of the form

- $i(a) \circ i(\tilde{a})$,
- $\sum_{(s,\tilde{s}) \in g} i(s \to \tilde{s}) \circ \sum_{(s,\tilde{s}) \in \tilde{g}} i(s \to \tilde{s})$, or
- $i(a) \geq 0$,

where $\circ \in \{<, \leq\}$. Because this is a homogeneous system of equations, i.e., if $(i^*(a))_{a \in \mathcal{A}}$ is a solution, then, for all $\lambda > 0$, we have that $(\lambda \cdot i^*(a))_{a \in \mathcal{A}}$ is also a solution, we can replace strict inequalities by inequalities that are not strict by adding $+1$ to the corresponding side of the inequality. Then, we consider a system of inequalities where $\circ$ is replaced by $\leq$ or $\leq 1+$. Finally, we arrived at a system of linear equations constructed from the Puiseux policy σ. Because this system of linear equations has a solution, it has a solution in the rational. Moreover, the numerators and denominators of a solution can be bounded by Cramer's rule so they use polynomial size. Multiplying the rational solution of this system to obtain an integer solution of the original set of (strict) inequalities we finally conclude the existence of rank policy witnesses. □

Lemma (Restated, Lemma 7)**.** *There exists a polynomial-time algorithm that, given a blind MDP and a rank policy, decides whether the rank policy is a witness of limit-sure reachability or not.*

Proof. Consider a blind MDP and a rank policy. The algorithm constructs the graph of communicating classes iteratively and checks whether the only absorbing communicating class reachable from the initial state s_0 is $\{\top\}$ or not. Concretely, the algorithm constructs this graph similar to the proof of Lemma 6 as follows.

- Initialization. Start by considering a communicating class for each singleton state.
- Adding edges. Given the currently considered communicating classes, add all edges given by the support of their exit distribution.
- Adding new communicating classes. Given the currently considered communicating classes, consider those that are not contained in another communicating class. With the support of their exit distribution, by Lemma 3, we find larger communicating classes if there are any.

By repeating the last two items until no other communicating class is found, we obtain the full graph of communicating classes. We show that this algorithm runs in polynomial time.

By Corollary 1, there are at most $(2|\mathcal{S}| - 1)$ communicating classes. There are two relevant operations for the algorithm that computes the graph of communicating classes. First, computing the support of the exit distribution of a communicating class. Second, checking whether a new communicating class should be added. We show how to perform these operations in polynomial time.

Fix a communicating class $\mathcal{B} \subsetneq \mathcal{S}$. We compute the support of its exit distributions, i.e., $\operatorname{supp}(\delta(\mathcal{B}, \cdot))$. Consider $\tilde{s} \in \mathcal{S} \setminus \mathcal{B}$. By Theorem 5 and Theorem 6, the state $\tilde{s}$ is in the support of the exit distribution of $\mathcal{B}$ if and only if there exists a state $s \in \mathcal{S}$ and an exit graph $g \in G_{\mathcal{B}}(s \to \tilde{s})$ whose weight coincides with the smallest weight among all exit graphs in $G_{\mathcal{B}}$, i.e., the weight of some exit graph is $G_{\mathcal{B}}^{\min}$. We determine this in two steps. First, we compute the weight of some exit graph in $G_{\mathcal{B}}^{\min}$. Second, we compute the minimum weight over all exit graphs in $\cup_{s \in \mathcal{S}} G_{\mathcal{B}}^{\min}(s \to \tilde{s})$. Comparing these quantities we determine whether $\tilde{s}$ is in the support of the exit distribution of $\mathcal{B}$. For the first step, collapse all states in $(\mathcal{S} \setminus \mathcal{B})$ into a single state and compute a minimal directed spanning tree where the weight of an edge is given by the leading power in the Puiseux power expansion of the corresponding transition. A directed spanning tree in this collapsed graph corresponds to an exit graph in the Puiseux Markov chain, and their weights coincide. By Gabow et al. [1986], this computation takes polynomial time in $|\mathcal{S}|$. This determines the weight of some exit graph in $G_{\mathcal{B}}^{\min}$. For the second step, for all $s \in \mathcal{S}$, we proceed similarly, i.e., we collapse all states in $\{s\} \cup (\mathcal{S} \setminus \mathcal{B})$ into a single state and compute a minimal directed spanning tree. Then, we add the smallest weight among the transitions from s to $\tilde{s}$. The result corresponds to the weight of an exit graph in $G_{\mathcal{B}}^{\min}$ containing the edge $(s, \tilde{s})$. By Gabow et al. [1986], this computation takes at polynomial time in $|\mathcal{S}|$, and we repeat it at most $|\mathcal{B}| \leq |\mathcal{S}|$ times. Taking the minimum over all computed weights while varying $s \in \mathcal{B}$, we deduce the minimum weight of all graphs in $\cup_{s \in \mathcal{S}} G_{\mathcal{B}}^{\min}(s \to \tilde{s})$. Recall that this weight coincides with the one of an exit graph in $G_{\mathcal{B}}^{\min}$ if and only if $\tilde{s}$ is in the support of the exit distribution. Therefore, comparing the weights obtained in the first and second steps we decide whether $\tilde{s}$ is in the support or not.

Given all communicating classes computed so far and the support of their exit distributions, we check whether we can add another communicating class. By the characterization in Lemma 3, this corresponds to computing bottom strongly connected components in a directed graph of at most $(2|\mathcal{S}| - 1)$ vertices, which takes linear time by Dijkstra [1976], and is done at most $(|\mathcal{S}| - 1)$ times.

Finally, given the full graph of communicating classes induced by the rank policy, we check whether the policy is a witness of limit-sure reachability. By the characterization in Lemma 5, we run a depth-first search starting at $\{s_0\}$ and check whether the only reachable absorbing communicating class from s_0 is $\{\top\}$. We conclude that checking whether the rank policy is a witness of limit-sure reachability or not takes polynomial time. □

B PROOFS OF SECTION 3.3

In this section, we provide the detailed proofs of results in Section 3.3.

Proposition (Restated, Proposition 1)**.** *For all constants $m \geq 0$, the problem of determining whether a POMDP P with reachability objective is limit-sure winning under memory m policies is NP-hard.*

In the rest of the section, for completeness, we give an explicit reduction from 3-SAT [Karp, 1972] that prove Proposition 1.

3-SAT. Consider Boolean variables $x_1, x_2, \ldots, x_n$ and clauses $C_1, C_2, \ldots, C_m$ where each clause is the disjunction of three literals from the set $\{x_1, x_2, \ldots, x_n\} \cup \{\neg x_1, \neg x_2, \ldots, \neg x_n\}$. The 3-SAT problem is determining if there is an assignment of the Boolean variables that satisfies all clauses.

Proof. Consider an instance of 3-SAT given by Boolean variables $x_1, x_2, \dots, x_n$ and clauses $C_1, C_2, \dots, C_m$. For $j \in [m]$, denote clause $C_j = \ell(j_1) \vee \ell(j_2) \vee \ell(j_3)$, where each literal $\ell \in \{x_1, x_2, \dots, x_n\} \cup \{\neg x_1, \neg x_2, \dots, \neg x_n\}$. We construct the POMDP given as follows.

- $\mathcal{S} \coloneqq \sqcup_{j \in [m]} \{\ell(j,1), \ell(j,2), \ell(j,3)\} \cup \{s_0, \top, \bot\}$, where $\ell(j,k)$ is a different state for each $j \in [m]$ and $k \in [3]$ that represents the k-th Boolean variable of the j-th clause;
- $\mathcal{A} \coloneqq \{t, f\}$ is the action set which represents assigning a truth value to a variable;
- $\delta \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function given by

$$\delta(s,a) = \begin{cases} \frac{1}{m} \sum\limits_{j \in [m]} \mathbb{1}[\ell(j,1)] & s = s_0 \\ \mathbb{1}[\top] & \begin{array}{l} a = t, \exists j \in [m], k \in [3] : \\ \quad s = \ell(j,k) = x_i \end{array} \\ \mathbb{1}[\top] & \begin{array}{l} a = f, \exists j \in [m], k \in [3] : \\ \quad s = \ell(j,k) = \neg x_i \end{array} \\ \mathbb{1}[\ell(j,k+1)] & \begin{array}{l} a = f, \exists j \in [m], k \in [2] : \\ \quad s = \ell(j,k) = x_i \end{array} \\ \mathbb{1}[\ell(j,k+1)] & \begin{array}{l} a = t, \exists j \in [m], k \in [2] : \\ \quad s = \ell(j,k) = \neg x_i \end{array} \\ \mathbb{1}[\bot] & \begin{array}{l} a = f, \exists j \in [m] : \\ \quad s = \ell(j,3) = x_i \end{array} \\ \mathbb{1}[\bot] & \begin{array}{l} a = t, \exists j \in [m] : \\ \quad s = \ell(j,3) = \neg x_i \end{array} \\ \mathbb{1}[s] & s \in \{\top, \bot\} \end{cases}$$

  In other words: s_0 moves to the first literal of each clause uniformly independent of the action; each literal moves to either $\top$ or the next literal in the clause; the terminal states $\top$ and $\bot$ are absorbing, i.e., for all actions $a \in \mathcal{A}$, we have that $\delta((s,a)) = \mathbb{1}[s]$, for $s \in \{\top, \bot\}$.
- $\mathcal{Z} \coloneqq \{x_i : i \in [n]\} \cup \{s_0, \top, \bot\}$ is the set of observations, one per Boolean variable;
- $o \colon \mathcal{S} \to \mathcal{Z}$ is the observation function that forces the controller to assign a consistent truth value to the literals and is given by

$$o(s) = \begin{cases} s & s \in \{s_0, \top, \bot\} \\ x_i & \begin{array}{l} \exists i \in [n], j \in [m], k \in [3] : \\ \quad s = \ell(j,k) \in \{x_i, \neg x_i\} \end{array} \end{cases}$$

- $s_0 \in \mathcal{S}$ is the initial state.

See Figure 3 for an illustration of this reduction. We show that this POMDP has reachability value 1 if and only if the 3-SAT instance is satisfiable.

Assume the 3-SAT instance is satisfiable with a valuation $v \colon \{x_i : i \in [n]\} \to \{t, f\}$. Consider the memoryless deterministic policy for the controller given by any extension of the valuation, for example assigning action t to states that do not represent a Boolean variable. In other words, consider $\sigma \colon \mathcal{Z} \to \mathcal{A}$ given by

$$\sigma(s) = \begin{cases} v(x_i) & \exists i \in [n] : s = x_i \\ t & s \in \{s_0, \top, \bot\} \end{cases}$$

We show that this policy guarantees a reachability probability of one, and therefore the POMDP has a reachability value of one.

From the initial state s_0, any action leads uniformely to the set $\{\ell(j,1) : j \in [m]\}$. Therefore, it is enough to show that starting from $\ell(j,1)$ we reach $\top$, for an arbitrary $j \in [m]$. Fix $j \in [m]$. Since the clause $C_j = \ell(j_1) \vee \ell(j_2) \vee \ell(j_3)$ is satisfied by the valuation v, we consider three cases depending on the first literal that evaluates to true.

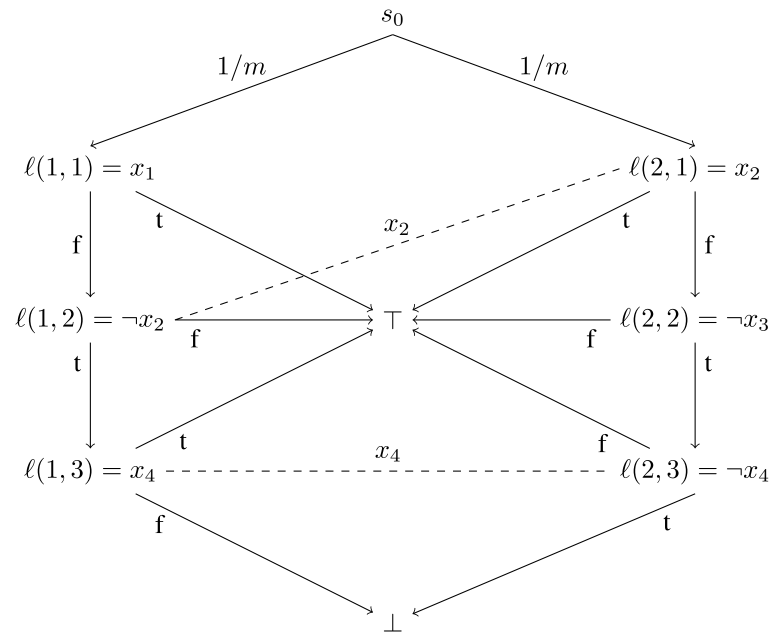

Figure 3: Example of the reduction from 3-SAT to the limit-sure winning reachability problem in POMDPs for the 3-SAT instance $(x_1 \vee \neg x_2 \vee x_4) \wedge (x_2 \vee \neg x_3 \vee \neg x_4)$.

1. Assume $v(\ell(j,1)) = t$ and fix $i \in [n]$ such that $\ell(j,1) \in \{x_i, \neg x_i\}$. Then, by the definition of σ and δ, we have that $\delta(\ell(j,1), \sigma(x_i)) = \mathbb{1}[\top]$. In other words, $\ell(j,1)$ reaches $\top$ in a single transition.
2. Assume $v(\ell(j,1)) = f$ and $v(\ell(j,2)) = t$ and fix $i_1, i_2 \in [n]$ such that $\ell(j,k) \in \{x_{i_k}, \neg x_{i_k}\}$ for $k \in [2]$. Then, by the definition of σ and δ, we have that $\delta(\ell(j,1), \sigma(x_{i_1})) = \mathbb{1}[\ell(j,2)]$ and $\delta(\ell(j,2), \sigma(x_{i_2})) = \mathbb{1}[\top]$. In other words, $\ell(j,1)$ reaches $\top$ after two transitions.
3. Assume $v(\ell(j,1)) = f$, $v(\ell(j,2)) = f$, and $v(\ell(j,3)) = t$ and fix $i_1, i_2, i_3 \in [n]$ such that $\ell(j,k) \in \{x_{i_k}, \neg x_{i_k}\}$ for $k \in [3]$. Then, by the definition of σ and δ, we have that $\delta(\ell(j,1), \sigma(x_{i_1})) = \mathbb{1}[\ell(j,2)]$, $\delta(\ell(j,2), \sigma(x_{i_2})) = \mathbb{1}[\ell(j,3)]$, and $\delta(\ell(j,3), \sigma(x_{i_3})) = \mathbb{1}[\top]$. In other words, $\ell(j,1)$ reaches $\top$ after three transitions.

Since in all cases, starting from $\ell(j,1)$ we reach $\top$, we have proven that σ guarantees a reachability value of one.

Assume the 3-SAT instance is not satisfiable, i.e., for all valuations $v \colon \{x_i : i \in [n]\} \to \{t, f\}$ there exists at least one clause that evaluates to false. We show that every deterministic memoryless policy leads to a reachability value strictly less than one and therefore the POMDP does not have a reachability value of one.

Note that the reachability value of our POMDP may consider only deterministic policies. This observation holds POMDPs with general policies [Feinberg, 1996, Venel and Ziliotto, 2016], and we argue that this holds for our POMDP even when considering only memoryless policies because there are no loops in the dynamic. Indeed, our POMDP can be seen as an extended-form game with one player. Moreover, the controller may remember all of their previous actions of the game since, during the process, no observation is presented twice before reaching the states $\top$ and $\perp$, where the outcome is determined. Therefore, Kuhn's theorem [Aumann, 1964, Section 5] applies and every memoryless policy induces the same distribution over outcomes that some distribution over deterministic policies. In other words, the reachability value of our POMDP may consider only deterministic policies.

Consider a deterministic memoryless policy $\sigma \colon \mathcal{Z} \to \mathcal{A}$. Define the valuation $v \colon \{x_i : i \in [n]\} \to \{t, f\}$ given by $v(x_i) = \sigma(x_i)$. Since the 3-SAT instance is not satisfiable, there exists a clause C_j with $j \in [m]$ such that $v(C_j) = f$. Therefore, under σ, starting from $\ell(j,1)$, the dynamic does not reach $\top$ by a similar argument as before. We conclude that the reachability probability starting from s_0 and following σ is at most $1 - 1/m$, and therefore the POMDP does not have value one. This concludes the proof of NP-hardness. □

C PROOFS OF SECTION 3.4

In this section, we provide the detailed proofs of results in Section 3.4.

Corollary (Restated, Corollary 2)**.** *The problem of determining whether a POMDP $P = (\mathcal{S}, \mathcal{A}, \delta, \mathcal{Z}, o, s_0)$ with reachability objective is limit-sure winning under constant memory policies is NP-complete.*

Proof. The NP-hardness follows from Proposition 1. The NP upper bound is obtained as follows. Consider an amount of memory $m \geq 1$. Note that an update function $\sigma_u \colon [m] \times \mathcal{Z} \times \mathcal{A} \to [m]$ is a finite object of polynomial size. Moreover, fixing an update function σ_u and considering the product POMDP with states $\mathcal{S} \times [m]$, memoryless policies in the product correspond to policies with memory m in the original POMDP. The formal definition of the product POMDP is $P_m \coloneqq \left(\mathcal{S} \times [m], \mathcal{A}, \tilde{\delta}, \mathcal{Z} \times [m], \tilde{o}, (s_0, 1)\right)$ where

- the observation function $\tilde{o}$ is defined as, for all $s \in \mathcal{S}$ and $\mu \in [m]$,

$$\tilde{o}\left((s, \mu)\right) \coloneqq (o(s), \mu)\,,$$

- the transition function $\tilde{\delta}$ is given by

$$\tilde{\delta}((s, \mu), a)((\tilde{s}, \tilde{\mu})) \coloneqq \begin{cases} \delta(s, a)(\tilde{s}) & \sigma_u(\mu, o(s), a) = \tilde{\mu} \\ 0 & \sim \end{cases}$$

Then, P_m is limit-sure winning under memoryless policies if and only if P is limit-sure winning under memory policies using memory amount m, which proves the NP upper bound. □

Corollary (Restated, Corollary 3)**.** *The problem of determining whether a POMDP P with parity objective is limit-sure winning under constant memory policies is NP-complete.*

Proof. The NP-hardness follows from Proposition 1. The NP upper bound is obtained as follows. First, consider memoryless policies because the proof for constant memory policies follows from the reduction described in the proof of Corollary 2.

Consider a POMDP that is limit-sure winning for the parity objective under memoryless policies. Then, by definition of the limit-sure winning property, for all $\varepsilon > 0$, there exists a policy σ_ε such that the probability of satisfying the parity condition is at least $1 - \varepsilon$ under this policy. Consider a sequence $(\varepsilon_n \coloneqq 1/n)_{n \geq 1}$ and a corresponding sequence of policies $(\sigma_n)_{n \geq 1} \subseteq \Delta(\mathcal{A})^{\mathcal{Z}}$. Since the set of all possible supports per observation is finite, up to taking a subsequence, we assume that, for all $z \in \mathcal{Z}$, the support of $\sigma_n(z)$ is invariant on n. For a Markov chain, the recurrent classes depend only on the support of the transition function. Therefore, the recurrent classes of the Markov chains induced by σ_n do not depend on n.

Note that, in a Markov chain, the parity condition is satisfied if and only if a good recurrent class is reached. Therefore, the probability of satisfying the parity objective in the POMDP under σ_n corresponds to the reachability probability to good recurrent classes (under σ_n). By guessing the support of a sequence of policies that are witness of the limit-sure property for the POMDP, we reduce the parity objective to the reachability objective of the corresponding good recurrent classes. □

Corollary (Restated, Corollary 4)**.** *The problem of determining whether a parametric Markov chain with reachability objective is limit-sure winning under constant memory policies is NP-complete.*

Proof. An extension of the quantitative reachability objective, denoted reach-avoid objective, was considered by [Junges et al., 2018]. They showed that the quantitative problem for POMDPs and parametric Markov chains (pMCs) for reach-avoid objectives are equivalent [Junges et al., 2018, Corollary 1]. In particular, they are equivalent for the reachability objective. Moreover, the reduction from a POMDP to a pMC for a quantitative objective does not depend on the quantitative threshold used in the objective. Therefore, the problem whether a POMDP is limit-sure winning reduces to the problem whether a pMC is limit-sure winning. The same occurs with the reduction from a pMC to a POMDP, establishing the equivalence of both problems. □