Revelations: A Decidable Class of POMDPs with Omega-Regular Objectives

Marius Belly[1], **Nathanaël Fijalkow**[1], **Hugo Gimbert**[1],
Florian Horn[2], **Guillermo A. Pérez**[3], **Pierre Vandenhove**[1*]

[1]CNRS, LaBRI, Université de Bordeaux, France
[2]CNRS, IRIF, Université de Paris, France
[3]University of Antwerp – Flanders Make, Antwerp, Belgium

Abstract

Partially observable Markov decision processes (POMDPs) form a prominent model for uncertainty in sequential decision making. We are interested in constructing algorithms with theoretical guarantees to determine whether the agent has a strategy ensuring a given specification with probability 1. This well-studied problem is known to be undecidable already for very simple omega-regular objectives, because of the difficulty of reasoning on uncertain events. We introduce a revelation mechanism which restricts information loss by requiring that almost surely the agent has eventually full information of the current state. Our main technical results are to construct exact algorithms for two classes of POMDPs called *weakly* and *strongly revealing*. Importantly, the decidable cases reduce to the analysis of a finite belief-support Markov decision process. This yields a conceptually simple and exact algorithm for a large class of POMDPs.

Code — https://github.com/gaperez64/pomdps-reveal
Extended version — https://arxiv.org/abs/2412.12063

1 Introduction

Partially observable Markov decision processes (POMDPs) form a prominent model for uncertainty in sequential decision making. They were defined in the 1960s (Åström 1965) for operations research and introduced in artificial intelligence by the seminal paper of Kaelbling, Littman, and Cassandra (1998). We consider POMDPs from a model-based point of view common in planning and in formal methods. Our goal is to construct exact (as opposed to approximate) algorithms that take as an input a complete description of the POMDP and construct a strategy ensuring a given specification. A long line of work has established that most formulations of this problem are undecidable. For instance, even in the extreme case where the agent has no information and the goal is to reach a target state with arbitrarily high probability, complex convergence phenomena occur, implying strong undecidability results (Madani, Hanks, and Condon 2003; Gimbert and Oualhadj 2010; Fijalkow 2017).

In this work, we are interested in constructing *almost-sure strategies*, meaning strategies ensuring their specifications with probability 1. We consider the class of omega-regular objectives (all expressible as *parity objectives*), which is a robust class including properties expressible in Linear Temporal Logic (Pnueli 1977; Giacomo and Vardi 2013). Determining whether there exists an almost-sure strategy against the subclass of CoBüchi objectives (requiring to avoid a target from some point onwards) is undecidable (Chatterjee, Chmelik, and Tracol 2016; Bertrand, Genest, and Gimbert 2017). There is a vast body of work towards approximate and practical solutions: for instance, using interpolation in the belief space (Lovejoy 1991), approximation of the value function (Hauskrecht 2000), or Monte Carlo tree search approaches (Silver and Veness 2010). This is orthogonal to the current paper since we focus on exact algorithms.

Our starting point is a simple approach to construct almost-sure strategies: from the POMDP, we build a Markov decision process (MDP) whose states are *supports of the beliefs* of the POMDP. In other words, we store information about which states we can be in, but abstract away the probabilities. The *belief-support MDP* serves as a finite abstraction of the POMDP; one could expect that there exists an almost-sure strategy in the POMDP if and only if there exists one in the corresponding belief-support MDP. Unfortunately, this abstraction is neither sound nor complete; we present a simple counterexample in Figure 1.

The fundamental question we ask is whether **there are natural sufficient conditions making the belief-support abstraction correct**. Conceptually, the failure of this abstraction is due to the accumulation of information loss over time.

We introduce a **revelation mechanism** which restricts information loss by requiring that, almost surely, the agent has eventually full information of the current state. Intuitively, by forbidding information loss from accumulating for an unbounded amount of time, the revelation mechanism removes the convergence issues leading to undecidability. Practically, we conjecture that revelation is a commonly occurring phenomenon in partial observability; a canonical example is systems with a small probability of resetting infinitely often, and where this reset is observable. We leave to future work to investigate this question further. Other approaches to restrict information loss have been proposed; we refer to the related works (Section 6) for an additional discussion.

*Authors are listed in alphabetical order.

Figure 1: We consider the POMDP on the LHS: there is a single signal s, so no information is ever given about the exact state we are in (a behavior the revelation mechanisms forbid!). Yet, almost surely, we reach q_1. The *priorities* indicated on states constitute a parity condition inducing the objective "eventually never visiting q_1", which clearly cannot be ensured almost surely. We represent the *belief-support MDP* on the RHS: the two states are $\{q_0\}$ and $\{q_0, q_1\}$, and only the state $\{q_0, q_1\}$ is visited infinitely often. To assign priorities to the states of this MDP, there are two natural candidates: "maximal priority semantics" and "minimal priority semantics", meaning that we assign either the maximal or minimal priority from the states in the belief support. In this figure, we use the maximal priority semantics: the priority of $\{q_0, q_1\}$ is thus 2, so the belief-support MDP is winning. This means that the analysis of the belief-support MDP is not sound in general. By tweaking the priorities in this example, one can show that both priority semantics are neither sound nor complete.

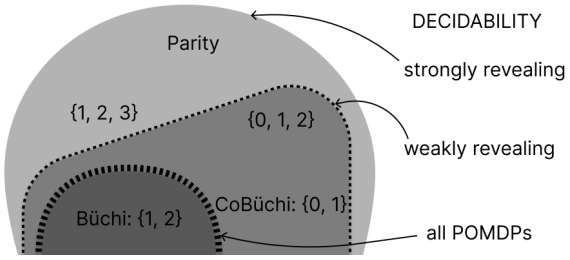

Figure 2: Summary of our results: decidable subclasses of the *parity* objective depending on the revelation mechanism.

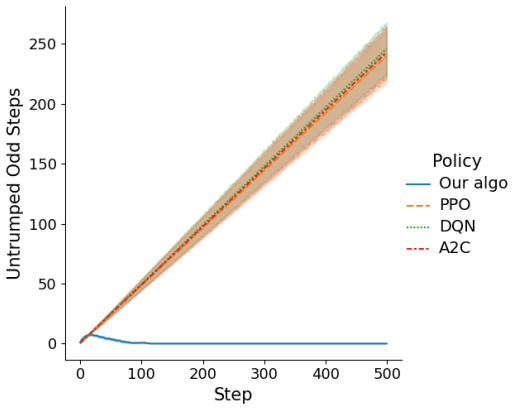

Figure 3: Omega-regular specifications have a natural interpretation in terms of *bad* events that must all be *trumped* by future *good* events. Along a simulation of the POMDP, one can keep track of the number of steps from the last bad event that has not yet been trumped (i.e., lower is better). Here, we depict this value, per step (from 1 to 500) over 500 simulations of a revealing version of the classical tiger POMDP (Cassandra, Kaelbling, and Littman 1994). A2C, DQN, and PPO are (`MlpPolicy`) strategies obtained from the *stable-baselines* library (Raffin et al. 2021), trained (for a total of 10k time steps) with default parameter values using a simple reward scheme: a good event yields a reward of 100; a bad one, -1. In the simulations, the trained models are queried for deterministic action predictions. The example used will be discussed in Section 5, Example 2.

Our contributions.

- We study two properties of POMDPs based on the revelation mechanism, called *weak* and *strong revelations*.
- We obtain decidability (and undecidability) results for both classes. Importantly, the decidable cases reduce to the analysis of the finite belief-support MDP. A summary of our contributions for POMDPs is provided in Figure 2. We also briefly consider the class of *two-player games of partial information*, to show that our revealing mechanisms do not suffice for decidability on this larger class.
- We provide a simple implementation of the algorithm as a proof of concept. We provide a comparison between our algorithm and off-the-shelf deep reinforcement learning (DRL) trained via an observation wrapper. As we will show in the paper, the MDP induced by the belief supports carries sufficient information to play in revealing POMDPs; hence, we used a wrapper implementing a subset construction on the fly to generate the current belief support, and focused on algorithms intended for MDPs. Spending moderate effort on reward engineering and hyperparameter tuning, we have been unable to match the performance of our algorithm (see Figure 3).

This yields a conceptually simple and exact algorithm for a large class of POMDPs. The importance of our results can be appreciated by the following remark: instead of a subclass of POMDPs, the revelation mechanism can be seen as new semantics for *all* POMDPs. In that sense, we obtain decidability results for an *optimistic* semantics of POMDPs which, to the best of our knowledge, has not been done before. We refer to Section 5.3 for more details on this point of view.

Extended version. Due to a lack of space, proofs are omitted from this version. They are in the appendix of the extended version (Belly et al. 2024), along with additional details and examples.

2 Preliminaries

A *(discrete) probability distribution* on a finite set X is a function $d\colon X \to [0,1]$ such that $\sum_{x\in X} d(x) = 1$. The set of all probability distributions on X is denoted $\mathcal{D}(X)$. The *support* $\mathsf{supp}(d)$ of a probability distribution d is the set

$\{x \in X \mid d(x) > 0\}$. We let $|X|$ denote the number of elements in a set X.

2.1 POMDPs

A *partially observable Markov decision process* (POMDP) is a tuple $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ such that Q is a finite set of *states*, Act is a finite set of *actions*, Sig is a finite state of *signals*, $\delta\colon Q \times \mathsf{Act} \to \mathcal{D}(\mathsf{Sig} \times Q)$ is the *transition function*, and $q_0 \in Q$ is an *initial state*.

A *play* of a POMDP $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ is an infinite sequence $\pi = q_0 a_1 s_1 q_1 a_2 s_2 \ldots \in (Q \cdot \mathsf{Act} \cdot \mathsf{Sig})^\omega$ such that, for all $i \geq 0$, $\delta(q_i, a_{i+1})(s_{i+1}, q_{i+1}) > 0$. A *history* h of a POMDP is a finite prefix of a play ending in a state (it is an element of $(Q \cdot \mathsf{Act} \cdot \mathsf{Sig})^* \cdot Q$). If $h = q_0 a_1 s_1 q_1 \ldots a_n s_n q_n$, we write $\mathsf{last}(h)$ for q_n. In practice, states are not fully observable; we define an *observable history* as the projection of a history to the subsequence in $(\mathsf{Act} \cdot \mathsf{Sig})^*$. We write $\mathsf{obs}(h)$ for the observable history derived from a history h, i.e., the same sequence with the states removed.

For q a state of a POMDP $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$, we define $\mathcal{P}^q$ to be the POMDP $(Q, \mathsf{Act}, \mathsf{Sig}, \delta, q)$ with only a change of initial state. We let $\beta_\mathcal{P} = \min\{\delta(q,a)(s,q') \mid q, q' \in Q,\ a \in \mathsf{Act},\ s \in \mathsf{Sig}, \text{ and } \delta(q,a)(s,q') > 0\}$ denote the least non-zero probability occurring in $\mathcal{P}$.

A *Markov decision process* (MDP) is a tuple $\mathcal{M} = (Q, \mathsf{Act}, \delta, q_0)$ where $\delta\colon Q \times \mathsf{Act} \to \mathcal{D}(Q)$. Formally, an MDP $\mathcal{M} = (Q, \mathsf{Act}, \delta, q_0)$ can be seen as a POMDP $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ such that $\mathsf{Sig} = \{s_q \mid q \in Q\}$ and for all $q, q', q'' \in Q$ and $a \in \mathsf{Act}$, $\delta(q,a)(s_{q''}, q') > 0$ if and only if $q' = q''$. In practice, it means that the last signal always uniquely determines the current state. MDPs have "complete observation", whereas POMDPs have "partial observation". For a POMDP $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$, we define the *underlying MDP of* $\mathcal{P}$ to be the MDP $(Q, \mathsf{Act}, \delta', q_0)$ with $\delta'(q,a)(q') = \sum_{s \in \mathsf{Sig}} \delta(q,a)(s,q')$.

Remark 1. *The observable information in POMDPs is here provided through* signals *that appear along transitions. This contrasts with state-based* observations *that partition the state space, which are also frequently used to model POMDPs. Both models are polynomially equivalent: a POMDP with observations can be transformed into an equivalent POMDP with signals on the same state space, while the converse requires an increase of the state space linear in* $|\mathsf{Sig}|$*. Both choices are convenient, but using signals make the definition of* strongly revealing *(Definition 2) more natural, which is why we opted for this convention.*

Strategies. Let $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ be a POMDP. An *(observation-based) strategy in* $\mathcal{P}$ is a function that makes decisions based on the current observable history, i.e., it is a function $\sigma\colon (\mathsf{Act} \cdot \mathsf{Sig})^* \to \mathcal{D}(\mathsf{Act})$. We can define strategies in MDPs similarly (i.e., assuming that Sig gives the information of the current state), but we assume for convenience that a strategy is a function $\sigma\colon (\mathsf{Act} \cdot Q)^* \to \mathcal{D}(\mathsf{Act})$ in this case. An observable history $a_1 s_1 \ldots a_n s_n$ is *consistent with a strategy* σ if for all $1 \leq i < n$, $\sigma(a_1 s_1 \ldots a_i s_i)(a_{i+1}) > 0$.

A strategy σ is *pure* if for all observable histories $h \in (\mathsf{Act} \cdot \mathsf{Sig})^*$, $\sigma(h)$ is a Dirac distribution; in other words, if σ is a function $(\mathsf{Act} \cdot \mathsf{Sig})^* \to \mathsf{Act}$. We let $\Sigma(\mathcal{P})$ denote the set of strategies in POMDP $\mathcal{P}$ and $\Sigma_\mathsf{P}(\mathcal{P})$ denote the set of pure strategies in $\mathcal{P}$.

For an MDP $\mathcal{M}$, a strategy σ in $\mathcal{M}$ is *memoryless* if its decisions are only based on the current state: i.e., if for all histories h_1, h_2, $\mathsf{last}(h_1) = \mathsf{last}(h_2)$ implies $\sigma(h_1) = \sigma(h_2)$. We only define the memoryless notion for MDPs.

Probability measure induced by a strategy. Let $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ be a POMDP. For a history h of $\mathcal{P}$, we define $\mathsf{Cyl}(h)$ (the *cylinder of* h) to be the set of all plays starting with h, i.e., $h(\mathsf{Act} \cdot \mathsf{Sig} \cdot Q)^\omega$. Given a strategy σ, we can define a probability measure $\mathbb{P}^\mathcal{P}_\sigma[\cdot]$ on infinite plays. This function is naturally defined over cylinders by induction. We define $\mathbb{P}^\mathcal{P}_\sigma[\mathsf{Cyl}(q_0)] = 1$, and $\mathbb{P}^\mathcal{P}_\sigma[\mathsf{Cyl}(q)] = 0$ for $q \in Q$, $q \neq q_0$. For a history $h = h'asq$, we define $\mathbb{P}^\mathcal{P}_\sigma[\mathsf{Cyl}(h)] = \mathbb{P}^\mathcal{P}_\sigma[\mathsf{Cyl}(h')] \cdot \sigma(\mathsf{obs}(h'))(a) \cdot \delta(\mathsf{last}(h'), a)(s, q)$. By Ionescu-Tulcea extension theorem (Klenke 2007), this function can be uniquely extended to a probability distribution $\mathbb{P}^\mathcal{P}_\sigma[\cdot]$ over the Borel sets of infinite plays induced by all cylinders.

We use this probability distribution to measure sets of infinite sequences in Q^ω, by associating a set $W \subseteq Q^\omega$ with the set $\bigcup_{q_0 q_1 \ldots \in W} q_0 \mathsf{ActSig} q_1 \mathsf{ActSig} q_2 \ldots \subseteq (Q \times \mathsf{Act} \times \mathsf{Sig})^\omega$. Similarly, we use this probability distribution to measure events based on signals, by associating a set $S \subseteq \mathsf{Sig}^\omega$ with the set $\bigcup_{s_1 s_2 \ldots \in S} Q \mathsf{Act} s_1 Q \mathsf{Act} s_2 Q \ldots \subseteq (Q \times \mathsf{Act} \times \mathsf{Sig})^\omega$.

Objectives. Let $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ be a POMDP. An *objective* $W \subseteq Q^\omega$ is a measurable set of infinite sequences of states. Note that observing an infinite sequence of signals (but not the states) may not always be sufficient to determine whether a play satisfies an objective.

Given a set $F \subseteq Q$, the *reachability objective* $\mathsf{Reach}(F) = \{q_0 q_1 \ldots \in Q^\omega \mid \exists i \geq 0, q_i \in F\}$ is the set of plays that visit a state in F at least once. For $k \in \mathbb{N}$, we write $\mathsf{Reach}^{\leq k}(F) = \{q_0 q_1 \ldots \in Q^\omega \mid \exists i, 0 \leq i \leq k, q_i \in F\}$ for the set of plays that reach F in at most k steps. Given a set $F \subseteq Q$, the *safety objective* $\mathsf{Safety}(F)$ is the set of plays that never visit any state in F.

Given a *priority function* $p\colon Q \to \{0, \ldots, d\}$ (where $d \in \mathbb{N}$), the *parity objective* $\mathsf{Parity}(p) = \{q_0 q_1 \ldots \in Q^\omega \mid \limsup_{i \geq 0} p(q_i) \text{ is even}\}$ is the set of infinite plays whose highest priority seen infinitely often is even. A *Büchi objective* is a parity objective $\mathsf{Parity}(p)$ such that $p\colon Q \to \{1, 2\}$, and a *CoBüchi objective* is a parity objective $\mathsf{Parity}(p)$ such that $p\colon Q \to \{0, 1\}$. For $Q' \subseteq Q$, we write $\mathsf{Büchi}(Q')$ for the set of infinite plays that visit Q' infinitely often. It is equal to $\mathsf{Parity}(p)$ for the priority function p such that $p(q) = 2$ if $q \in Q'$, and $p(q) = 1$ otherwise.

For an objective W, a strategy σ is *almost sure* if $\mathbb{P}^\mathcal{P}_\sigma[W] = 1$, and is *positively winning* if $\mathbb{P}^\mathcal{P}_\sigma[W] > 0$. We say that an objective W has *value* 1 in a POMDP $\mathcal{P}$ if $\sup_{\sigma \in \Sigma(\mathcal{P})} \mathbb{P}^\mathcal{P}_\sigma[W] = 1$.

2.2 Beliefs and Belief Supports

Let $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ be a POMDP. A *belief* $\mathfrak{b} \in \mathcal{D}(Q)$ is a probability distribution on Q. A *belief support* $b \in 2^Q \setminus \{\emptyset\}$ is the support of a belief. For brevity, we write $2^Q_\emptyset$ for $2^Q \setminus \{\emptyset\}$. At every step, beliefs and belief supports can be updated when playing an action and observing a signal.

We show how to do so for belief supports: we define a function $\mathcal{B}\colon 2^Q_\emptyset \times \mathsf{Act} \times \mathsf{Sig} \to 2^Q_\emptyset$ that updates the belief support. For $b \in 2^Q_\emptyset$, $a \in \mathsf{Act}$, $s \in \mathsf{Sig}$, we define $\mathcal{B}(b,a,s) = \{q' \in Q \mid \exists q \in b, \delta(q,a)(s,q') > 0\}$. We extend this function in a natural way to a function $\mathcal{B}^*\colon 2^Q_\emptyset \times (\mathsf{Act}\cdot\mathsf{Sig})^* \to 2^Q_\emptyset$. Objectives $\mathsf{Reach}(B)$ and $\mathsf{Büchi}(B)$ can be naturally extended to sets of belief supports $B \subseteq 2^Q_\emptyset$ (Belly et al. 2024, Appendix C).

Beliefs carry more information than belief supports, as they contain the exact probability of being in a particular state, while belief supports only contain the qualitative information of the possible current states. Observe that when the belief support is a singleton (i.e., $b = \{q\}$ for some $q \in Q$), knowing the precise belief does not yield more information than knowing the belief support, as all the probability mass is in one of the states. Our "revealing" restrictions on POMDPs defined later will exploit this fact.

3 The Belief-Support MDP

For a POMDP $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$, the *belief-support MDP of* $\mathcal{P}$ is the MDP $\mathcal{P}_\mathcal{B} = (2^Q_\emptyset, \mathsf{Act}, \delta_\mathcal{B}, \{q_0\})$ where for $b, b' \in 2^Q_\emptyset$ and $a \in \mathsf{Act}$, $\delta_\mathcal{B}(b,a)(b') > 0$ if and only if there is $s \in \mathsf{Sig}$ such that $\mathcal{B}(b,a,s) = b'$. We assume the distribution to be uniform over successors with positive probability.

We can show that for some simple objectives, the POMDP and its belief-support MDP behave in a similar way. For example, sets of belief supports that can be reached with a positive probability are the same in the POMDP and its belief-support MDP; if a set of belief supports is reachable almost surely in the POMDP, it is also the case in the belief-support MDP (formal statements and proofs in (Belly et al. 2024, Appendix C)).

There is a natural way to lift a strategy in the belief-support MDP to a strategy in the POMDP. We define a notation to go from a sequence of signals to the induced sequence of belief supports. Let $h = a_1 s_1 \dots a_n s_n \in (\mathsf{Act}\cdot\mathsf{Sig})^*$ be a possible observable history in $\mathcal{P}$. For $1 \le i \le n$, let $b_i = \mathcal{B}^*(\{q_0\}, a_1 s_1 \dots a_i s_i)$ be the belief support after i steps. We define B_h to be the history $a_1 b_1 \dots a_n b_n$ of $\mathcal{P}_\mathcal{B}$. Let $\sigma_\mathcal{B} \in \Sigma(\mathcal{P}_\mathcal{B})$ be a strategy in the belief-support MDP of a POMDP $\mathcal{P}$. We define a strategy $\widehat{\sigma_\mathcal{B}}$ in $\mathcal{P}$ derived from the strategy $\sigma_\mathcal{B}$: for $h \in (\mathsf{Act}\cdot\mathsf{Sig})^*$, we fix $\widehat{\sigma_\mathcal{B}}(h) = \sigma_\mathcal{B}(B_h)$.

4 Weakly Revealing POMDPs

We define here our first *revealing* property for POMDPs, which requires that, infinitely often and almost surely, the current state can be deduced by looking at the previous sequence of signals. Formally, we write $B^\mathcal{P}_{\mathsf{sing}} = \{\{q\} \mid q \in Q\}$ for the set of singleton belief supports of a POMDP $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$. An observable history $h \in (\mathsf{Act}\cdot\mathsf{Sig})^*$ such that $\mathcal{B}^*(\{q_0\}, h) \in B^\mathcal{P}_{\mathsf{sing}}$ is called a *revelation*.

Definition 1 (Weakly revealing)**.** *A POMDP $\mathcal{P}$ is* weakly revealing *if, for all strategies $\sigma \in \Sigma(\mathcal{P})$, we have $\mathbb{P}^\mathcal{P}_\sigma\big[\mathsf{Büchi}(B^\mathcal{P}_{\mathsf{sing}})\big] = 1$; i.e., for all strategies, infinitely many revelations occur almost surely.*

In particular, POMDPs that "reset" infinitely often, and whose reset can be observed with a dedicated signal, are weakly revealing. We will use one such example in Figure 4.

One can give probabilistic bounds on the occurrence of a revelation for a weakly revealing POMDP (Belly et al. 2024, Appendix D): starting from any reachable belief, a revelation occurs within $2^{|Q|}-1$ steps with probability at least $\beta_\mathcal{P}^{2^{|Q|}-1}$. The bound is asymptotically tight: there is a weakly revealing POMDP with $n+2$ states, 1 action, and n signals where we need at least $2^n - 1$ steps before observing a revelation with positive probability (Belly et al. 2024, Appendix E).

The decidability of the weakly revealing property itself is discussed in (Belly et al. 2024, Section 4.4 and Appendix F). A straightforward argument shows that is decidable in 2-EXPTIME. First, extend the POMPD with the information of the current belief support, which makes the state space exponential-sized. Then, check whether there exists a strategy that sees only finitely many singleton sets with positive probability, which is a positive CoBüchi objective and is itself decidable in EXPTIME (Chatterjee, Chmelik, and Tracol 2016).

4.1 Soundness of the Belief-Support MDP

In this section, we show that, for *weakly revealing* POMDPs, the existence of an almost-sure strategy in the belief-support MDP (with an adequate priority function) implies the existence of an almost-sure strategy in the POMDP.

For the priority function of the belief-support MDP, we consider the "maximal priority" semantics. Formally, let $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ be a POMDP, and $\mathcal{P}_\mathcal{B}$ be its belief-support MDP. Let $p\colon Q \to \{0, \dots, n\}$ be a priority function on $\mathcal{P}$, inducing the objective $\mathsf{Parity}(p)$. We extend this function to the belief-support MDP: for $b \in 2^Q_\emptyset$, we define

$$p_\mathcal{B}(b) = \max\{p(q) \mid q \in b\}.$$

Without any assumption, the belief-support MDP may be unsound, already for Büchi objectives; there may be an almost-sure strategy in the belief-support MDP, but not in the POMDP. An example illustrating this was given in Figure 1. Surprisingly, it is sound for CoBüchi objectives without any assumption (Belly et al. 2024, Appendix E). Using "$\max$" (and not "$\min$") turns out to be the right choice in our setting. Intuitively, under the right revealing assumptions and the right strategies, if a belief support is visited infinitely often, then all its states will be visited infinitely often, so the maximal priority of the belief support is the one that matters given the parity objective. Without any assumption, both $\max$ and $\min$ are unsound and incomplete in general.

Under the weakly revealing semantics, almost-sure strategies of the belief-support MDP carry over to the POMDP for all parity objectives. In other words, the analysis of the belief-support MDP is sound. We recall that pure memoryless strategies suffice to reach the optimal value for parity objectives in MDPs (Chatterjee and Henzinger 2012). The proof is in (Belly et al. 2024, Appendix E).

Proposition 1. *Let $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ be a weakly revealing POMDP with priority function p, and let $\mathcal{P}_\mathcal{B}$ be its belief-support MDP with priority function $p_\mathcal{B}$. Assume*

there is an almost-sure strategy $\sigma_{\mathcal{B}}$ for $\mathsf{Parity}(p_{\mathcal{B}})$ in $\mathcal{P}_{\mathcal{B}}$; by (Chatterjee and Henzinger 2012), we may assume $\sigma_{\mathcal{B}}$ to be pure and memoryless. Then, $\widehat{\sigma_{\mathcal{B}}}$ is an almost-sure strategy for $\mathsf{Parity}(p)$ in $\mathcal{P}$.

4.2 Decidability for Priorities 0, 1, and 2

We show that the existence of an almost-sure strategy in a weakly revealing POMDP implies the existence of an almost-sure strategy in its belief-support MDP when priorities are in $\{0,1,2\}$. This provides a converse to Proposition 1 when priorities are restricted to $\{0,1,2\}$. We will see that this is not the case for priorities in $\{1,2,3\}$ in the next section; this result is therefore optimal w.r.t. the priority used. We emphasize that parity objectives with priorities $\{0,1,2\}$ encompass both Büchi and CoBüchi objectives. This result is false without the weakly revealing assumption; see the simple POMDP in Figure 1. The proof is in (Belly et al. 2024, Appendix E).

Proposition 2. *Let $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ be a weakly revealing POMDP with priority function p with values in $\{0,1,2\}$. Let $\mathcal{P}_{\mathcal{B}}$ be its belief-support MDP with priority function $p_{\mathcal{B}}$. If there is an almost-sure strategy for $\mathsf{Parity}(p)$ in $\mathcal{P}$, then there is an almost-sure strategy for $\mathsf{Parity}(p_{\mathcal{B}})$ in $\mathcal{P}_{\mathcal{B}}$.*

From the above, we deduce a complexity upper bound; a matching lower bound is discussed in Section 5.

Theorem 1. *The existence of an almost-sure strategy for parity objectives with priorities in $\{0,1,2\}$ in weakly revealing POMDPs is EXPTIME-complete.*

Proof. The EXPTIME algorithm is a consequence of the results from this section: by Proposition 1 (soundness of the belief-support MDP) and Proposition 2 (completeness), we reduce the problem to the existence of an almost-sure strategy for a parity objective with priorities in $\{0,1,2\}$ in an MDP of size exponential in $|Q|$. The existence of an almost-sure strategy for parity objectives is decidable in polynomial time in MDPs (Baier and Katoen 2008, Theorem 10.127). Proposition 1 also constructs an almost-sure strategy in $\mathcal{P}$.

The EXPTIME-hardness follows from Proposition 4 below, already for CoBüchi objectives and for the more restricted class of *strongly revealing* POMDPs. □

Remark 2. *The algorithm also gives an upper bound on the size of the strategies for parity objectives with priorities in $\{0,1,2\}$ in weakly revealing POMDPs. As we reduce to the analysis of an exponential-size MDP and that memoryless strategies suffice for parity objectives in MDPs, given Proposition 1, it means that a strategy of exponential size suffices in the POMDP. We can also prove an exponential lower bound (Belly et al. 2024, Appendix E).*

4.3 Undecidability for Priorities 1, 2, and 3

The previous section suggests that analyzing the belief-support MDP is a sound and complete approach for weakly revealing POMDPs with parity objectives with priorities in $\{0,1,2\}$. One may wonder whether it is complete for any priority function. Unfortunately, this fails to hold in general, already for priority functions taking values in $\{1,2,3\}$. We discuss one such example below.

Example 1. *Consider the POMDP $\mathcal{P}$ in Figure 4. This POMDP is weakly revealing, as state q_0 is visited infinitely often for any strategy and is revealed through signal s_0. The only choice in this POMDP is in states q_1 and q_1': whether to play a and move to q_0 or $\{q_1, q_1'\}$, or to play c and go to q_2 or q_3. Observe that when the game starts in q_0, the only reachable belief supports are $\{q_0\}$, $\{q_1, q_1'\}$, and $\{q_2, q_3\}$, which all have a maximal odd priority. Hence, the belief-support MDP with priority function $p_{\mathcal{B}}$ trivially has no almost-sure (and even positively) winning strategy. However, we show that there is an almost-sure strategy in $\mathcal{P}$.*

The only way to win in this POMDP is to visit q_2 infinitely often while visiting q_3 only finitely often. To do so, observe that when a is played multiple times in a row and only receives signal s_1, the probability to be in q_1' becomes arbitrarily close to 1. Formally, if σ_a is the strategy that only plays a, we have that for $n > 0$,

$$\mathbb{P}^{\mathcal{P}}_{\sigma_a}[Q^n q_1' \mid (s_1)^n] = 1 - \mathbb{P}^{\mathcal{P}}_{\sigma_a}[q_0 (q_1)^n \mid (s_1)^n] = 1 - \frac{1}{2^n}.$$

For $n > 0$, let σ_n be the strategy that plays only a until s_1 has been seen n times in a row*, and when that is the case, plays c. Let us divide a play in this POMDP into rounds 1, 2,...; every time we go back to q_0 after visiting q_2 or q_3, we move to the next round. Consider the strategy that plays σ_n in round n. This strategy ensures that infinitely many rounds happen, because at each round n, it will eventually succeed in seeing n occurrences of s_1 in a row. At each round n, c is eventually played with probability 1. By the above equation, q_3 is seen with probability $\frac{1}{2^n}$ and q_2 is seen with probability $1 - \frac{1}{2^n}$. State q_2 is clearly seen infinitely often almost surely. However, the probability that q_3 is never seen anymore after round n is equal to $\prod_{i=n}^{\infty}(1 - \frac{1}{2^i})$, which is positive and increases as n grows to ∞. We deduce that the probability that q_3 is seen at most finitely often is 1.*

Generalizing the above example, we show that if we allow p to take values in $\{1,2,3\}$, the existence of almost-sure strategies in weakly revealing POMDPs is undecidable. We provide here a proof sketch; a full proof is in (Belly et al. 2024, Appendix E).

Theorem 2. *The existence of an almost-sure strategy in weakly revealing POMDPs with a parity objective with priorities in $\{1,2,3\}$ is undecidable. The same holds for the existence of a positively winning strategy.*

Our proof uses a reduction from the value-1 problem in *probabilistic automata*. A *probabilistic automaton* (Rabin 1963) is a tuple $\mathcal{A} = (Q, \mathsf{Act}, \delta, q_0)$. One can define their semantics through POMDPs: they behave like POMDPs in which we assume that the signals bring no information (Sig is a singleton). No useful information is provided by the signals along a play (beyond the number of steps played); pure strategies therefore correspond to words on alphabet Act.

Intuitively, the proof expands on the POMDP in Figure 4 by replacing states q_1, q_1' by a copy of a probabilistic automaton $\mathcal{A}$: the transition from q_0 goes to the initial state of $\mathcal{A}$, and playing c goes to q_2 if the current state is a final state

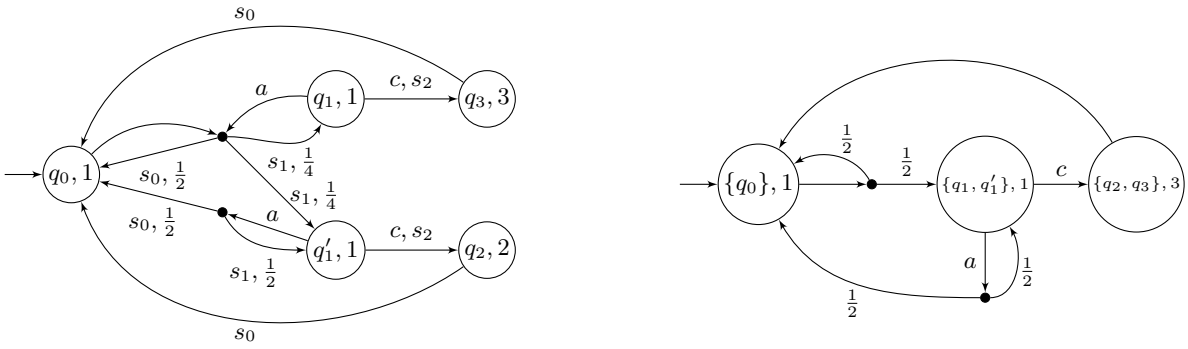

Figure 4: The POMDP $\mathcal{P}$ from Example 1 (depicted on the left) with an almost-sure strategy, but whose belief-support MDP (depicted on the right) has no winning strategy. Notation q, k inside a circle depicts a state q with priority k. Transitions from states involving a bullet $\bullet$ indicate a probabilistic transition. In POMDPs, we always write the signals along transitions. Actions are omitted when they all induce the same transition from a given state, and probabilities equal to 1 are omitted.

of $\mathcal{A}$, and to q_3 otherwise. We keep a positive probability to go back to q_0 at any point to make it weakly revealing. The idea of playing n times a in a row in the example is replaced by a (possible) sequence of words that have a probability arbitrarily close to 1 to reach a final state. One can show that there is an almost-sure strategy in this POMDP if and only if $\mathcal{A}$ has value 1 w.r.t. its final states.

5 Strongly Revealing POMDPs

In this section, we introduce *strongly revealing POMDPs*, a stronger property entailing that infinitely many revelations occur in a POMDP almost surely. We show that the existence of almost-sure strategies is decidable for strongly revealing POMDPs with arbitrary parity objectives.

We define a notion of *revealing signals*: for q a state of a POMDP $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$, we define $\mathsf{Revealing}(q) = \{s \in \mathsf{Sig} \mid \forall r, r' \in Q, r' \neq q \implies \delta(r, a)(s, r') = 0\}$ to be the set of signals that indicate surely that the next state is q. For convenience, we define $\mathsf{Succ}(q, a) = \{q' \in Q \mid \exists s \in \mathsf{Sig}, \delta(q, a)(s, q') > 0\}$ and $\mathsf{Succ}(q, a, s) = \{q' \in Q \mid \delta(q, a)(s, q') > 0\}$.

Definition 2. *POMDP $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ is* strongly revealing *if any transition between two states for a given action in the underlying MDP of $\mathcal{P}$ can also happen with a revealing signal. Formally, $\mathcal{P}$ is strongly revealing if for all $q, q' \in Q$ and $a \in \mathsf{Act}$, if $q' \in \mathsf{Succ}(q, a)$, then there is $s \in \mathsf{Revealing}(q')$ such that $q' \in \mathsf{Succ}(q, a, s)$.*

Under this definition, the set of belief supports $B^{\mathcal{P}}_{\mathsf{sing}}$ is visited infinitely often from the initial state for any given strategy, so a strongly revealing POMDP is in particular weakly revealing. Observe that the weakly revealing POMDP from Figure 4 is not strongly revealing: for instance, $q_1' \in \mathsf{Succ}(q_1, a)$, but there is no revealing signal that could for sure reveal q_1' after q_1. The strongly revealing property can be decided in polynomial time in the size of a POMDP by simply analyzing every transition.

Example 2. *We give an example of a strongly revealing POMDP inspired from the* tiger *of (Cassandra, Kaelbling, and Littman 1994), depicted in Figure 5. This example was used in Figure 3 in the introduction; the code to generate it in our tool is provided in (Belly et al. 2024, Appendix A).*

In the tiger environment, an agent has to open the left or the right door, with action a_{L} or a_{R}, respectively. One of them has a (deadly) tiger behind it. Fortunately, the agent can choose to wait and listen (action $a_?$) to help its decision. Listening results in a signal that is biased towards the reality, i.e., the signal can be s_{L} or s_{R} and the former is more likely if the tiger really is on the left, and vice versa.

We present our version of the tiger environment in which listening guarantees one will eventually discern behind which door there is a tiger. This is achieved by adding new revealing signals $a_{\mathsf{L}!}$ or $a_{\mathsf{R}!}$ which, importantly, can only be obtained when the tiger is on the left or on the right, respectively. To keep things interesting, these signals can only be obtained with a small probability (yet, them being there already ensures that the POMDP is strongly revealing*). We also add signals for death ($s_\bot$) and victory ($s_\top$), which are missing from the original tiger environment.*

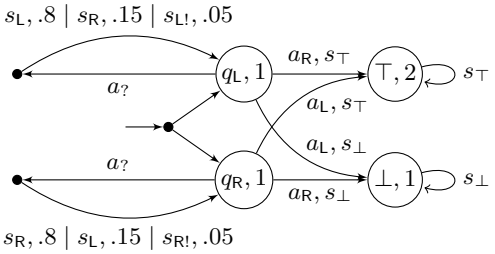

Figure 5: Strongly revealing *tiger* (Example 2).

5.1 Decidability of Parity with Strong Revelations

The soundness of the analysis of the belief-support MDP for strongly revealing POMDPs follows from Proposition 1; it remains to show completeness (proofs for this section in (Belly et al. 2024, Appendix G)).

Proposition 3. *Let $\mathcal{P} = (Q, \mathsf{Act}, \mathsf{Sig}, \delta, q_0)$ be a strongly revealing POMDP with a priority function p, and let $\mathcal{P}_{\mathcal{B}}$ be*

its belief-support MDP with priority function $p_{\mathcal{B}}$. If there is an almost-sure strategy for $\mathsf{Parity}(p)$ *in* $\mathcal{P}$*, then there is an almost-sure strategy for* $\mathsf{Parity}(p_{\mathcal{B}})$ *in* $\mathcal{P}_{\mathcal{B}}$.

We also show a complexity lower bound. The lower bound holds for CoBüchi in strongly revealing POMDPs; as strongly revealing POMDPs are a subclass of weakly ones, the hardness follows for weakly revealing POMDPs.

Proposition 4. *The following problem is EXPTIME-hard: given a strongly revealing POMDP with a CoBüchi objective, decide whether there is an almost-sure strategy.*

We obtain as before the decidability of the problem by reducing to the analysis of the belief-support MDP. The proof is similar to the one of Theorem 1, simply replacing the use of Proposition 2 by Proposition 3.

Theorem 3. *The existence of an almost-sure strategy for parity objectives in strongly revealing POMDPs is EXPTIME-complete.*

5.2 Undecidability of Strongly Revealing Games

We discuss here whether the revealing semantics helps in *zero-sum games* of partial information with revealing semantics. In general, such games with CoBüchi objectives are undecidable (they encompass POMDPs) while Büchi games are decidable for almost-sure strategies (Bertrand, Genest, and Gimbert 2017). We obtained a negative result: the existence of an almost-sure strategy in CoBüchi *games* with partial information is undecidable, even when satisfying a natural extension of the strongly revealing property. The model is defined formally in (Belly et al. 2024, Appendix G), along with an undecidability proof.

5.3 Optimistic Semantics for POMDPs

In our revealing definitions, we adopted the point of view of considering *subclasses* of POMDPs. A limitation of this point of view is that our results say nothing about POMDPs which are not strongly (nor weakly) revealing. We argue that another fruitful formulation of our results concerns the class of *all* POMDPs, by defining alternative, *revealing* semantics.

Consider a POMDP $\mathcal{P}$. Let us define the extended POMDP $\mathcal{P}_{\mathsf{sr}}$ such that, at each transition, there is a small probability of revealing which state we reach after firing this action, using additional signals s_q, one for each state q of $\mathcal{P}$.

Theorem 4. *For any POMDP* $\mathcal{P}$*,* $\mathcal{P}_{\mathsf{sr}}$ *is strongly revealing. Moreover, if there is no almost-sure strategy ensuring an omega-regular objective in* $\mathcal{P}_{\mathsf{sr}}$ *(which is decidable by Theorem 3), then there is no almost-sure strategy ensuring the same objective in* $\mathcal{P}$.

The contrapositive is easily proved: any almost-sure strategy of $\mathcal{P}$ can be lifted to an almost-sure strategy of $\mathcal{P}_{\mathsf{sr}}$. This property justifies the term "optimistic semantics". Note that the converse implication cannot hold (as POMDPs with omega-regular objectives are undecidable).

6 Related Works

We discuss additional references where a restriction is set to stochastic systems to make them decidable.

The closest idea to our revelations that we know of is in (Berwanger and Mathew 2017), defining a class of partial-information multi-player games with *sure* (not just almost-sure) revelations; from any point in the game, a "revelation" occurs surely within a bounded number of steps. This is a yet stronger kind of revelation mechanism under which even parity *games* are decidable.

The *decisiveness property* (Abdulla, Ben Henda, and Mayr 2007; Bertrand et al. 2020) is a useful property to decide reachability properties in infinite stochastic systems (without decision-making). Decisiveness is implied by the existence of a *finite attractor*; there is such an attractor in weakly revealing POMDPs once we fix a finite-memory strategy (as in Proposition 1).

Another path to decidability and strong guarantees is to restrict strategies, such as studying "memoryless" (Vlassis, Littman, and Barber 2012) or finite-memory (Chatterjee, Chmelik, and Tracol 2016; Andriushchenko et al. 2022) strategies in POMDPs. In our paper, the strategies we consider only use finite memory, as they are memoryless strategies on the belief-support MDP (in our case, they are even shown to be optimal among all strategies under the right assumptions). The sufficiency of belief-support-based strategies in POMDPs, which was known for almost-sure reachability (Baier, Größer, and Bertrand 2012), was also exploited to craft efficient algorithms in (Junges, Jansen, and Seshia 2021); such an approach could speed up our algorithms.

In a quantitative setting, the idea of having actions with some cost that reveal the current state or decrease the uncertainty has appeared multiple times in the literature. Such an idea appeared in 2011 (Bertrand and Genest 2011) for POMDPs with quantitative reachability objectives. Recently, *active-measuring POMDPs*, with a similar mechanism, have been considered in the online planning community (Bellinger et al. 2021; Krale, Simão, and Jansen 2023). Despite a different setting (online planning vs. model checking), it carries an intuition similar to our work: precise states can be known, which helps find good strategies.

Also in online planning, the article (Liu et al. 2022) considers a subclass of POMDPs restricting information loss that make *reinforcement learning* sample efficient.

7 Perspectives

We presented classes of POMDPs for which many natural objectives become decidable, and showed that these lie close to undecidability frontiers (priorities $\{0,1,2\}$ vs. $\{1,2,3\}$, POMDPs vs. games).

Due to their intrinsic undecidability, POMDPs are not often studied through the prism of exact algorithms. We believe there is a lot to gain by understanding more closely (i) the *structural properties* of POMDPs that make them decidable for classes of objectives (such as weak/strong revelations), and (ii) the conditions that make *simple strategies* (such as belief-support-based strategies) sufficient. Our article is a new step towards these goals. On a more specific note, an interesting step for (i) could involve framing the exact complexity of the existence of strategies for simple objectives involving beliefs.

Acknowledgments

This work was partially supported by the SAIF project, funded by the "France 2030" government investment plan managed by the French National Research Agency, under the reference ANR-23-PEIA-0006. Pierre Vandenhove was funded by ANR project G4S (ANR-21-CE48-0010-01). This work was sparked by discussions with Guillaume Vigeral and Bruno Ziliotto, following a talk on a related model (Vigeral and Ziliotto 2022).

References

Abdulla, P. A.; Ben Henda, N.; and Mayr, R. 2007. Decisive Markov Chains. *Log. Methods Comput. Sci.*, 3(4).

Andriushchenko, R.; Ceska, M.; Junges, S.; and Katoen, J. 2022. Inductive synthesis of finite-state controllers for POMDPs. In Cussens, J.; and Zhang, K., eds., *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, 85–95. PMLR.

Åström, K. J. 1965. Optimal Control of Markov Processes with Incomplete State Information I. *Journal of Mathematical Analysis and Applications*, 10: 174–205.

Baier, C.; Größer, M.; and Bertrand, N. 2012. Probabilistic ω-automata. *J. ACM*, 59(1): 1:1–1:52.

Baier, C.; and Katoen, J. 2008. *Principles of model checking*. MIT Press. ISBN 978-0-262-02649-9.

Bellinger, C.; Coles, R.; Crowley, M.; and Tamblyn, I. 2021. Active Measure Reinforcement Learning for Observation Cost Minimization. In Antonie, L.; and Zadeh, P. M., eds., *Canadian Conference on Artificial Intelligence*. Canadian Artificial Intelligence Association.

Belly, M.; Fijalkow, N.; Gimbert, H.; Horn, F.; Pérez, G. A.; and Vandenhove, P. 2024. Revelations: A Decidable Class of POMDPs with Omega-Regular Objectives. *CoRR*, abs/2412.12063.

Bertrand, N.; Bouyer, P.; Brihaye, T.; and Fournier, P. 2020. Taming denumerable Markov decision processes with decisiveness. *CoRR*, abs/2008.10426.

Bertrand, N.; and Genest, B. 2011. Minimal Disclosure in Partially Observable Markov Decision Processes. In Chakraborty, S.; and Kumar, A., eds., *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2011, December 12–14, 2011, Mumbai, India*, volume 13 of *LIPIcs*, 411–422. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Bertrand, N.; Genest, B.; and Gimbert, H. 2017. Qualitative Determinacy and Decidability of Stochastic Games with Signals. *J. ACM*, 64(5): 33:1–33:48.

Berwanger, D.; and Mathew, A. B. 2017. Infinite games with finite knowledge gaps. *Information and Computation*, 254: 217–237.

Cassandra, A. R.; Kaelbling, L. P.; and Littman, M. L. 1994. Acting Optimally in Partially Observable Stochastic Domains. In Hayes-Roth, B.; and Korf, R. E., eds., *Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, USA, July 31 - August 4, 1994, Volume 2*, 1023–1028. AAAI Press / The MIT Press.

Chatterjee, K.; Chmelik, M.; and Tracol, M. 2016. What is decidable about partially observable Markov decision processes with ω-regular objectives. *Journal of Computer and System Sciences*, 82(5): 878–911.

Chatterjee, K.; and Henzinger, T. A. 2012. A survey of stochastic ω-regular games. *Journal of Computer and System Sciences*, 78(2): 394–413.

Fijalkow, N. 2017. Undecidability results for probabilistic automata. *ACM SIGLOG News*, 4(4): 10–17.

Giacomo, G. D.; and Vardi, M. Y. 2013. Linear Temporal Logic and Linear Dynamic Logic on Finite Traces. In Rossi, F., ed., *International Joint Conference on Artificial Intelligence*, IJCAI'13, 854–860. IJCAI/AAAI.

Gimbert, H.; and Oualhadj, Y. 2010. Probabilistic Automata on Finite Words: Decidable and Undecidable Problems. In Abramsky, S.; Gavoille, C.; Kirchner, C.; auf der Heide, F. M.; and Spirakis, P. G., eds., *International Colloquium on Automata, Languages and Programming, ICALP*, volume 6199 of *Lecture Notes in Computer Science*, 527–538. Springer.

Hauskrecht, M. 2000. Value-Function Approximations for Partially Observable Markov Decision Processes. *Journal of Artificial Intelligence Research*, 13: 33–94.

Junges, S.; Jansen, N.; and Seshia, S. A. 2021. Enforcing Almost-Sure Reachability in POMDPs. In Silva, A.; and Leino, K. R. M., eds., *Computer Aided Verification – 33rd International Conference, CAV 2021, Virtual Event, July 20-23, 2021, Proceedings, Part II*, volume 12760 of *Lecture Notes in Computer Science*, 602–625. Springer.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1): 99–134.

Klenke, A. 2007. *Probability Theory: A Comprehensive Course*. Springer.

Krale, M.; Simão, T. D.; and Jansen, N. 2023. Act-Then-Measure: Reinforcement Learning for Partially Observable Environments with Active Measuring. In Koenig, S.; Stern, R.; and Vallati, M., eds., *Proceedings of the Thirty-Third International Conference on Automated Planning and Scheduling, Prague, Czech Republic, July 8–13, 2023*, 212–220. AAAI Press.

Liu, Q.; Chung, A.; Szepesvári, C.; and Jin, C. 2022. When Is Partially Observable Reinforcement Learning Not Scary? In Loh, P.; and Raginsky, M., eds., *Conference on Learning Theory, 2–5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, 5175–5220. PMLR.

Lovejoy, W. S. 1991. Computationally Feasible Bounds for Partially Observed Markov Decision Processes. *Operations Research*, 39(1): 162–175.

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

Pnueli, A. 1977. The temporal logic of programs. In *Symposium on Foundations of Computer Science*, SFCS'77.

Rabin, M. O. 1963. Probabilistic Automata. *Inf. Control.*, 6(3): 230–245.

Raffin, A.; Hill, A.; Gleave, A.; Kanervisto, A.; Ernestus, M.; and Dormann, N. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*, 22(268): 1–8.

Silver, D.; and Veness, J. 2010. Monte-Carlo Planning in Large POMDPs. In *Conference on Neural Information Processing Systems, NIPS*, 2164–2172. Curran Associates, Inc.

Vigeral, G.; and Ziliotto, B. 2022. Zero-sum stochastic games with intermittent observation of the state. Communication at the "Current Trends in Graph and Stochastic Games" workshop (GAMENET'22).

Vlassis, N.; Littman, M. L.; and Barber, D. 2012. On the Computational Complexity of Stochastic Controller Optimization in POMDPs. *ACM Trans. Comput. Theory*, 4(4): 12:1–12:8.