Revealing POMDPs: Qualitative and Quantitative Analysis for Parity Objectives

Ali Asadi[1], Krishnendu Chatterjee[1], David Lurie[2], and Raimundo Saona[3]

[1]Institute of Science and Technology Austria
[2]Paris Dauphine University, PSL Research University, Paris, France and NyxAir, Paris, France
[3]London School of Ecnonomics and Political Science, London, United Kingdom
{*ali.asadi, krishnendu.chatterjee*}*@ista.ac.at, david.lurie@dauphine.eu, raimundo.saona@gmail.com*

December 9, 2025

Abstract

Partially observable Markov decision processes (POMDPs) are a central model for uncertainty in sequential decision making. The most basic objective is the reachability objective, where a target set must be eventually visited, and the more general parity objectives can model all ω-regular specifications. For such objectives, the computational analysis problems are the following: (a) qualitative analysis that asks whether the objective can be satisfied with probability 1 (almost-sure winning) or probability arbitrarily close to 1 (limit-sure winning); and (b) quantitative analysis that asks for the approximation of the optimal probability of satisfying the objective. For general POMDPs, almost-sure analysis for reachability objectives is EXPTIME-complete, but limit-sure and quantitative analyses for reachability objectives are undecidable; almost-sure, limit-sure, and quantitative analyses for parity objectives are all undecidable. A special class of POMDPs, called revealing POMDPs, has been studied recently in several works, and for this subclass the almost-sure analysis for parity objectives was shown to be EXPTIME-complete. In this work, we show that for revealing POMDPs the limit-sure analysis for parity objectives is EXPTIME-complete, and even the quantitative analysis for parity objectives can be achieved in EXPTIME.

1 Introduction

POMDPs *Partially observable Markov decision processes* (POMDPs) model sequential decision-making under uncertainty [8, 33, 27]. At each step, the environment is in some hidden state. A controller interacts with it by choosing actions. The chosen action and the current state determine a probability distribution over the subsequent state. The controller cannot observe the state directly, but observes a signal that discloses only partial information of the state. POMDPs generalize classic models: *Markov decision processes* (MDPs), in which the state is fully observed [37], and *blind MDPs*, in which no state information is observed and are equivalent to Probabilistic Finite Automata [38, 34].

Objectives The controller aims to maximize an *objective function*, which formally captures the desired behaviors of the model. Two main classes of objectives are typically considered: *logical objectives*, e.g., reachability and LTL (linear-temporal logic), and *quantitative objectives*, e.g., discounted sum and limit-average; see [37, 6, 25] for details. This work focuses on logical objectives.

The most basic example of logical objectives is *reachability*, i.e., given a set of target states, the objective requires that some target state is visited at least once. A more general class of logical objectives is *parity objectives*, which assign to each state a non-negative integer, called *priority*. The objective is satisfied if the smallest priority appearing infinitely often is even. Parity objectives express all commonly used temporal objectives such as liveness, and are a canonical form to express all ω-regular objectives [40] and LTL objectives [6]. Hence, the study of POMDPs with parity objectives is a fundamental theoretical problem.

Applications POMDPs have been applied across diverse fields, including computational biology [23] and reinforcement learning [28]. In particular, POMDPs with logical objectives have proven useful in many application areas such as probabilistic planning [10]; randomized embedded scheduler [22]; randomized distributed algorithms [36]; probabilistic specification languages [5]; and robot planning [30, 11].

Computational Analysis Questions A policy determines the choice of actions by the controller. The value corresponds to the maximum probability that the controller can guarantee to satisfy an objective. The main computational analysis of POMDPs with reachability and parity objectives considers the following two problems.

(a) *Qualitative analysis* has two variants: the *almost-sure winning* asks whether there exists a policy that satisfies the objective with probability 1; and the *limit-sure winning* asks whether the objective can be satisfied with probability arbitrarily close to 1.

(b) *Quantitative analysis* asks to approximate the maximum or optimal probability with which the objective can be satisfied, up to a given additive error.

General Undecidability Most results for POMDPs and the computational analysis are negative (undecidability results). Indeed, the limit-sure analysis for reachability objectives is undecidable [26, 16], which extends to general parity objectives. The almost-sure analysis for reachability objectives is EXPTIME-complete [4, 14], but is undecidable for parity objectives even for two priorities (namely coBüchi objectives) [4, 13]. The quantitative analysis for reachability objectives is undecidable [32], which extends to general parity objectives. Given this wide range of results about non-existence of algorithms a natural direction to explore is the existence of subclasses for which the computational problems are decidable.

Revealing POMDPs We study *revealing POMDPs*, a special subclass of POMDPs where each visited state is announced to the controller with a positive probability. Consequently, the controller's uncertainty about the state, i.e. the probability distribution over states (the *belief*), occasionally collapses into a Dirac distribution on the announced state. Revealing POMDPs have been previously studied in several works, including [21, 7, 3], which present motivation and applications for this model.

Problems	Objectives	
	Reachability	Parity
Almost-sure	EXP-complete	Undecidable
Limit-sure	Undecidable	Undecidable
Quantitative	Undecidable	Undecidable

Table 1: Computational complexity for POMDPs with reachability and parity objectives.

Problems	Objectives	
	Reachability	Parity
Almost-sure	EXP	EXP-complete
Limit-sure	**EXP**	**EXP-complete**
Quantitative	**EXP**	**EXP**

Table 2: Computational complexity for revealing POMDPs with reachability and parity objectives. Our contributions are marked in bold.

Previous Results and Open Questions A key result for revealing POMDPs shows that the almost-sure analysis for parity objectives is EXPTIME-complete [7]. However, the limit-sure and quantitative analysis problems for reachability and parity objectives remained open for revealing POMDPs.

Our Contributions We address the open questions for revealing POMDPs. Our main contributions are the following. For revealing POMDPs with parity objectives, we show that

- The limit-sure analysis coincides with almost-sure analysis, and consequently is EXPTIME-complete.
- The quantitative analysis can be achieved in EXPTIME, i.e., in the same complexity as for qualitative analysis.

The results for POMDPs and revealing POMDPs are summarized in Table 1 and Table 2, respectively.

Technical Contributions A closely related work established that almost-sure analysis for parity objectives in revealing POMDPs is EXPTIME-complete [7]. Additionally, [12] previously showed that almost-sure analysis for parity objectives in general POMDPs under finite-memory policies is EXPTIME-complete. Therefore, the EXPTIME-completeness result naturally follows once finite memory policies are proven sufficient for almost-sure analysis. However, limit-sure analysis and quantitative analysis for POMDPs remain undecidable in general, even under finite memory policies [18]. This highlights a sharp contrast with the almost-sure analysis and gives rise to new technical challenges.

First, we consider revealing POMDPs with the belief-reachability objectives. Given a set of target states, the belief-reachability objectives consider the probability of eventually observing such states. We prove that the quantitative analysis of belief-reachability in revealing POMDPs can be achieved in EXPTIME. The argument proceeds in two steps: (i) we prove that this objective can be approximated by their finite-horizon counterparts; and (ii) we show that this finite-horizon objective can be approximated by a point-based approximation.

Then, we consider revealing POMDPs with the parity objectives. We show that the value for parity objectives corresponds to the value for belief-reachability objectives to almost-sure winning states. Finally, we prove the existence of optimal policies for parity objectives in revealing POMDPs.

Proofs and details omitted due to space restrictions are provided in the Appendix.

Related Works The study of subclasses of POMDPs with tractable algorithms is a broad topic with many directions. For example, subclasses of POMDPs for qualitative analysis have been explored in [24, 19]; and for quantitative analysis various subclasses have also been studied such as with ergodicity condition [17]; multiple environments only [41, 15]; or in online learning [31, 20]. Our work focuses on such a subclass, namely revealing POMDPs, which has been studied in the literature.

Avrachenkov, Dhiman, and Kavitha [3] studied strongly-connected revealing POMDPs and restricting attention to the set of belief-stationary policies. They proved that there is an optimal policy within this set, whose induced belief dynamics are contracting and reach a positive recurrent class of beliefs in finite time. They also provided an infinite-dimensional linear program that characterizes the optimal belief-stationary policy. They considered an application of strongly-connected revealing blind MDPs. [21] introduced the model of revealing blind MDPs. They consider the discounted objective and present algorithms based on finite-horizon approximation to compute the discounted value. Our work addresses the open computational analysis questions for revealing POMDPs.

2 Preliminaries

Notation For a positive integer n the set $\{1, 2, \ldots, n\}$ is denoted by $[n]$. Sets and correspondences are denoted by calligraphic letters, e.g., $\mathcal{S}, \mathcal{A}, \mathcal{Z}$. Elements of these sets are denoted by lowercase letters, e.g., s, a, z. Random elements with values in these sets are denoted by uppercase letters, e.g., S, A, Z. The set of probability measures over a finite set $\mathcal{S}$ is denoted by $\Delta(\mathcal{S})$. The Dirac measure at some element $s \in \mathcal{S}$ is denoted by $\mathbb{1}[s]$. The support of a probability measure $b \in \Delta(\mathcal{S})$ is denoted by $\operatorname{supp}(b)$.

Definition 1 (POMDP)**.** *A POMDP is a tuple $P = (\mathcal{S}, \mathcal{A}, \mathcal{Z}, \delta, b_0)$, where:*

- *$\mathcal{S}$ is a finite set of states;*
- *$\mathcal{A}$ is a finite set of actions;*
- *$\mathcal{Z}$ is a finite set of signals;*
- *$\delta \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S} \times \mathcal{Z})$ is a probabilistic transition function that, given a state s and an action a, returns the probability distribution over the successor states and signal;*
- *$b_0 \in \Delta(\mathcal{S})$ is the initial belief over the states.*

Markov Decision Processes (MDPs) are POMDPs in which the signal corresponds to the state [37]. Formally, an MDP is denoted by $M = (\mathcal{S}, \mathcal{A}, \delta, b_0)$ with transition function $\delta \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$.

Dynamic Given a POMDP, a controller knows all defining parameters. At the beginning, nature draws a state $S_0 \sim b_0$, which is not informed to the controller. Then, at each step $t \geq 0$, the controller chooses an action $A_t \in \mathcal{A}$, possibly at random. In response, nature draws the next state and signal $(S_{t+1}, Z_{t+1}) \sim \delta(S_t, A_t)$. The signal Z_{t+1} is revealed to the controller,

but the state S_{t+1} is not announced. A play (or a path) in the POMDP is an infinite sequence $\rho = (s_0, a_0, z_1, s_1, a_1, z_2, s_2, a_2, \ldots)$ of states, actions, and signals such that, for all $t \geq 0$, we have $\delta(s_t, a_t)(s_{t+1}, z_{t+1}) > 0$. The set of all plays is denoted by Ω.

Belief At each step $t \geq 0$, the controller's (random) belief about the current state can be computed using Bayes' rule and is denoted by $B_t \in \Delta(\mathcal{S})$. In the cases where the controller knows the exact state the controller's belief is a Dirac measure $\mathbb{1}[s]$ at some state $s \in \mathcal{S}$.

Definition 2 (Revealing POMDP)**.** *A POMDP is revealing if, each time a state is visited, the state is also announced to the controller with positive probability. Formally, for each state there is a designated signal, i.e., $\mathcal{S} \subseteq \mathcal{Z}$, and, for all states $s, s' \in \mathcal{S}$ and actions $a \in \mathcal{A}$,*

$$\sum_{z \in \mathcal{Z}} \delta(s,a)(s',z) > 0 \implies \sum_{\widetilde{s} \in \mathcal{S}} \delta(s,a)(\widetilde{s},s') = \delta(s,a)(s',s') > 0\,.$$

Revealing POMDPs coincide with the class of strongly revealing defined in [7]. Indeed, if the controller observes a signal $s \in \mathcal{Z}$, then their next belief is $\mathbb{1}[s]$, which corresponds to a revelation of the state.

Remark 1 (Signals)**.** *The signaling structure of POMDPs is modeled in different ways in the literature. We comment on two general cases.*

- *The controller may receive a set of signals at each step with a transition function $\delta\colon \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S} \times 2^{\mathcal{Z}})$. In that case, we consider each set of signals as one signal $\widetilde{\mathcal{Z}} \coloneqq 2^{\mathcal{Z}}$ and reduce to our model. Even with exponentially many signals encoded in a polynomial size input, our* EXPTIME *upper bound holds.*
- *In general POMDPs, it is enough to consider transition functions that pair each state with only one signal, known as deterministic observation, see [12, Remark 4]. For revealing POMDPs, deterministic observations do not capture the general case because they correspond to MDPs.*

Policies A *policy* defines how a controller selects actions based on all the information available up to a given step. Formally, a (history-dependent randomized) policy is a function $\sigma\colon (\mathcal{A}\times\mathcal{Z})^* \to \Delta(\mathcal{A})$. The set of all policies is denoted by Σ. A policy is pure if it prescribes deterministic actions, i.e., it corresponds to a function $\sigma\colon (\mathcal{A} \times \mathcal{Z})^* \to \mathcal{A}$.

Probability Measures For a finite prefix of a play, an element in $(\mathcal{S} \times \mathcal{A})^*$, its cone is the set of plays with it as their prefix. Given a policy σ and an initial belief b_0, the unique probability measure over Borel sets of infinite plays obtained given σ is denoted by $\mathbb{P}^{\sigma}_{b_0}(\cdot)$, which is defined by Carathéodory's extension theorem by extending the natural definition over cones of plays [9].

Objectives An objective in a POMDP is a Borel set of plays $\Phi \subseteq \Omega$ in the Cantor topology on Ω [29]. We consider objectives in the first $2\frac{1}{2}$ levels of the Borel hierarchy, including the parity objective which expresses all ω-regular objectives [40]. Denote the state at time t by s_t and set of states that occur infinitely often in a play ρ by $\mathcal{I}(\rho) \coloneqq \{s \in \mathcal{S} : \forall t \geq 0\ \exists \widetilde{t} \geq t \quad s = s_{\widetilde{t}} \in \rho\}$. We consider the following objectives.

- *Reachability:* Given a set $\mathcal{X} \subseteq \mathcal{S}$ of target states, the reachability objective requires that a target state is visited at least once. Formally, the reachability objective is

$$\text{Reach}(\mathcal{X}) \coloneqq \{\rho \in \Omega : \exists t \geq 0 \quad s_t \in \mathcal{X}\}.$$

- *Parity:* Given a $d \geq 0$ and a function $\text{pri}\colon \mathcal{S} \to \{0, 1, \ldots, d\}$ of priorities, the parity objective requires that the smallest priority that appears infinitely often is even. Formally,

$$\text{Parity} \coloneqq \{\rho \in \Omega : \min\{\text{pri}(s) : s \in \mathcal{I}(\rho)\} \text{ is even}\}.$$

Value The value of an objective in a POMDP is the maximum probability a controller can guarantee to satisfy the objective. Formally, given an objective $\Phi \subseteq \Omega$, the value is a function of the initial belief $\text{val}_\Phi\colon \Delta(\mathcal{S}) \to [0, 1]$ defined by $\text{val}_\Phi(b) \coloneqq \sup_{\sigma \in \Sigma} \mathbb{P}_b^\sigma(\rho \in \Phi)$. Denote the reachability and parity values by $\text{val}_{\text{R}(\mathcal{X})}$ and val_{P}, respectively. We may omit $\mathcal{X}$ in $\text{val}_{\text{R}(\mathcal{X})}$ if it is clear from the context.

Approximately Optimal Policies Given $\varepsilon \geq 0$ and an objective $\Phi \subseteq \Omega$, a policy $\sigma \in \Sigma$ is ε-optimal if it guarantees the value up to an additive error ε, i.e., if $\mathbb{P}_b^\sigma(\rho \in \Phi) \geq \text{val}_\Phi(b) - \varepsilon$. In particular, we call 0-optimal policies simply optimal.

Computational Analysis The computational analysis problems for POMDPs with an objective Φ are:

- *Almost-sure* analysis asks whether the objective can be satisfied with probability 1, i.e., $\exists \sigma \in \Sigma$ such that $\mathbb{P}_{b_0}^\sigma(\rho \in \Phi) = 1$.
- *Limit-sure* analysis asks whether the objective can be satisfied with probability arbitrarily close to 1, i.e., $\forall \varepsilon > 0\ \exists \sigma \in \Sigma$ such that $\mathbb{P}_{b_0}^\sigma(\rho \in \Phi) \geq 1 - \varepsilon$ or equivalently whether $\text{val}_\Phi(b_0) = 1$.
- *Quantitative* analysis asks to compute an approximation of the optimal value, i.e., for all $\varepsilon > 0$, provide $v \in [0, 1]$ such that $|v - \text{val}_\Phi(b_0)| \leq \varepsilon$.

3 Overview

We present an overview of our approach and the results.

3.1 Overview of Approach

To solve the qualitative and quantitative analysis of parity objectives, we introduce a new objective called belief-reachability and show that the parity value coincides with the belief-reachability value to the set of almost-sure winning parity states.

Belief-Reachability Given a set of target states $\mathcal{X} \subseteq \mathcal{S}$, the belief-reachability objective requires that a target state is visited and the controller has complete knowledge of the state at that step at least once. Formally, the set of Dirac beliefs on $\mathcal{X} \subseteq \mathcal{S}$ is denoted by $\mathcal{D}_\mathcal{X} \coloneqq \{b \in \Delta(\mathcal{S}) : \exists s \in \mathcal{X} \text{ such that } b = \mathbb{1}[s]\}$. Then, the belief-reachability objective is $\text{Belief-Reach}(\mathcal{X}) \coloneqq \{\rho \in \Omega : \exists t \geq 0 \quad B_t(\rho) \in \mathcal{D}_\mathcal{X}\}$. Denote the belief-reachability value by $\text{val}_{\text{BR}(\mathcal{X})}$. We may omit $\mathcal{X}$ in $\text{val}_{\text{BR}(\mathcal{X})}$ if it is clear from the context.

Remark 2 (Generality of belief-reachability)**.** *In general POMDPs with reachability objectives, without loss of generality, target states can be considered absorbing, i.e., the dynamic remains in the same state once a target state is visited. Furthermore, one can reduce to the case where there is only one target state. These simplifications affect the dynamic, but neither optimal policies nor the reachability value. Belief-reachability generalizes reachability objectives as follows. Given a POMDP P with an absorbing target state s^*, consider a copy of P but add a new action and two absorbing states $\top$ and $\bot$. After playing the new action, the state moves to $\top$ if the state was in s^*, and to $\bot$ otherwise. Moreover, the controller is announced which of these states is reached. Then, the belief-reachability value to $\top$ coincides with the original reachability value. Indeed, for an approximately optimal policy of P, play the new action after sufficiently many steps to obtain approximately the same value in the belief-reachability objective.*

3.2 Overview of Results

Results for Belief-Reachability Objectives Our main results for belief-reachability objectives are the following, which are proved in Section 4.

Theorem 1. *Quantitative analysis for belief-reachability objectives for revealing POMDPs is in* EXPTIME*.*

By Remark 2, reachability objectives reduce to belief-reachability objectives. Therefore, they have the following consequence.

Corollary 2. *Quantitative analysis for reachability objectives for revealing POMDPs is in* EXPTIME*.*

Results for Parity Objectives Our main results for parity objectives are the following, which are proved in Section 5.

Theorem 3. *Quantitative analysis for parity objectives for revealing POMDPs is in* EXPTIME*.*

Theorem 3 generalizes Corollary 2 to parity objectives, and follows from a reduction of parity objectives to belief-reachability to a set that can be computed in EXPTIME, see Lemma 15.

Theorem 4. *Optimal policies exist for parity objectives for revealing POMDPs.*

The following results follow from [7, Theorem 3] and Theorem 4.

Corollary 5. *For revealing POMDPs with parity objectives, limit-sure and almost-sure winning coincide, and limit-sure analysis is in* EXPTIME-complete*.*

4 Belief-Reachability Objectives

In this section, we prove Theorem 1, i.e., that the quantitative analysis for belief-reachability objectives is in EXPTIME. We proceed in five steps:

- We introduce reliable actions, which preserve the current belief-reachability value. Proposition 6 proves that every belief has a reliable action.
- Lemma 7 shows that stopping policies that play only reliable actions are optimal.
- Lemma 8 shows that playing reliable actions uniformly at random is stopping.

- Lemma 10 upper bounds the horizon needed to approximate the belief-reachability value, using stopping optimal policies.
- Lemma 11 presents a point-based dynamic programming algorithm to approximate the finite-horizon belief-reachability value.

Definition 3 (Reliable Action)**.** *Consider a POMDP with belief-reachability objectives. An action $a \in \mathcal{A}$ is* reliable *if it preserves the belief-reachability value. Formally, for each belief $b \in \Delta(\mathcal{S})$, define the set of reliable actions $\mathcal{R}(b)$ by*

$$\mathcal{R}(b) \coloneqq \left\{a \in \mathcal{A} : \mathbb{E}_b^a(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_1)) = \mathrm{val}_{\mathrm{BR}(\mathcal{X})}(b)\right\} .$$

Because the set of actions $\mathcal{A}$ is finite, we have the following result.

Proposition 6. *The set of reliable actions is nonempty.*

Definition 4 (Terminal State)**.** *Consider a POMDP and a set of target states $\mathcal{X} \subseteq \mathcal{S}$. A state $s \in \mathcal{S}$ is called* terminal *(for $\mathcal{X}$) if $s \in \mathcal{X}$ or if $\mathcal{X}$ cannot be reached from s with positive probability, i.e., $\sup_{\sigma\in\Sigma} \mathbb{P}^\sigma_{\mathbb{1}_{[s]}}(\exists t \geq 0 \quad \mathrm{supp}(B_t) \cap \mathcal{X} \neq \emptyset) = 0$. The set of terminal states is denoted by $\mathcal{T}$.*

Definition 5 (Stopping Policy)**.** *Consider a POMDP, a set of target states $\mathcal{X} \subseteq \mathcal{S}$, and a corresponding set of terminal states $\mathcal{T} \subseteq \mathcal{S}$. For $n \in \mathbb{N}$ and $q > 0$, a policy $\sigma \in \Sigma$ is called* (n, q)-stopping *(for $\mathcal{X}$) if within n steps the belief collapses to a Dirac mass on a terminal state with probability at least q. Formally, from every initial belief $b \in \Delta(\mathcal{S})$,*

$$\mathbb{P}^\sigma_b(\exists t \leq n \quad B_t \in \mathcal{D}_\mathcal{T}) \geq q\,.$$

We say that a policy is stopping if it is stopping for some n and $q > 0$.

Similar to [2, Lemma 5] that considers finite duration games, it is easy to see that, if a policy is stopping and uses only reliable actions, then it is optimal.

Lemma 7. *Consider a POMDP with belief-reachability objectives. If a policy is stopping and uses only reliable actions, then it is optimal.*

Proof Sketch. Consider a stopping policy $\sigma \in \Sigma$ which uses only reliable actions. Because every action chosen by σ is reliable, the belief-reachability values $\mathrm{val}_{\mathrm{BR}}(B_t)$ forms a martingale, i.e., $\mathrm{val}_{\mathrm{BR}}(b_0) = \mathbb{E}^\sigma_{b_0}\big(\mathrm{val}_{\mathrm{BR}}(B_t)\big)$. Moreover, since the policy σ is stopping, it reaches a Dirac belief on a terminal state in finite time with probability 1 and the hitting time τ of $\mathcal{D}_\mathcal{T}$ is almost-surely finite. Therefore, we have $\mathrm{val}_{\mathrm{BR}}(b_0) = \mathbb{E}^\sigma_{b_0}\big(\mathrm{val}_{\mathrm{BR}}(B_\tau)\big)$. □

Minimal Positive Transition Denote $\delta_{\min}$ the minimum non-zero probability in the transition function, i.e.,

$$\delta_{\min} \coloneqq \min\Big\{\delta(s,a)(s',z) \;:\; \forall s,\, s' \in \mathcal{S},\, a \in \mathcal{A},\, z \in \mathcal{Z} \quad \delta(s,a)(s',z) > 0\Big\}.$$

Lemma 8. *Consider a revealing POMDP with belief-reachability objectives. The policy σ that plays reliable actions uniformly at random is (n, q)-stopping with parameters $n \coloneqq |\mathcal{S}| + 2$ and $q \coloneqq \delta^2_{\min}\,(\delta_{\min}/|\mathcal{A}|)^{|\mathcal{S}|}$.*

Proof Sketch. By Proposition 6, every belief has a reliable action, so the policy σ is well-defined. Build the layered sets $\mathcal{S}_0 \coloneqq T$, $\mathcal{S}_{t+1} \coloneqq \{s : \exists a \in \mathcal{R}(\mathbb{1}[s])$ with a positive-probability move to $\mathcal{S}_t\}$. The sequence covers all states after at most $|\mathcal{S}|$ iterations, i.e., $\mathcal{S}_{|\mathcal{S}|} = \mathcal{S}$, because, if there were states outside, then they require to use unreliable actions to get to terminal states, which lowers their value and forms a contradiction. Under the policy σ that plays every reliable action uniformly, (a) the first "reveal" of the Dirac belief happens with probability $\delta_{\min}$; (b) each step toward the next layer then succeeds with probability at least $\delta_{\min}/|\mathcal{A}|$; and (c) the final "reveal" of the Dirac belief happens with probability $\delta_{\min}$. Hence, within $n \coloneqq |\mathcal{S}| + 2$ steps, a Dirac belief is reached on a terminal state with probability $q \coloneqq \delta_{\min}^2 \left(\delta_{\min}/|\mathcal{A}|\right)^{|\mathcal{S}|}$, which proves that σ is (n, q)-stopping. □

We deduce directly from Lemma 7 and Lemma 8 the next result.

Corollary 9. *Every revealing POMDP with belief-reachability objectives has an optimal policy.*

We turn to the quantitative analysis for the belief-reachability value. Note that the existence of optimal stopping policies implies that this value can be approximated using a finite horizon.

Lemma 10. *Consider a POMDP with belief-reachability objectives and an (n, q)-stopping optimal policy σ. For all $\varepsilon > 0$, the finite horizon $T \coloneqq n \left\lceil \frac{\log(\varepsilon)}{\log(1-q)} \right\rceil$ is such that*

$$\mathbb{P}^{\sigma}_{b_0}(\exists t \leq T \quad B_t \in \mathcal{D}_{\mathcal{T}}) \geq \mathrm{val}_{\mathrm{BR}}(b_0) - \varepsilon\,.$$

Proof Sketch. Split the play into blocks of n steps. Because the optimal policy σ is (n, q)-stopping, each block reaches a Dirac belief on a terminal state with probability at least q. Thus for the hitting time τ of $\mathcal{D}_{\mathcal{T}}$ can be bounded by $\mathbb{P}^{\sigma}_{b_0}(\tau > kn) \leq (1-q)^k$. Therefore, τ has a geometric tail and is almost surely finite. Choose $T = n \left\lceil \dfrac{\log(\varepsilon)}{\log(1-q)} \right\rceil$ so that $\mathbb{P}^{\sigma}_{b_0}(\tau > T) \leq \varepsilon$, showing that a horizon of T steps suffices to approximate the belief-reachability value. □

Reduction to Finite Horizon To solve the quantitative analysis for revealing POMDPs with belief-reachability objectives, it suffices to compute the finite horizon value for a horizon that is exponentially large on the input size. Indeed, by Lemma 8, there exists an (n, q)-stopping optimal policy with $n = |\mathcal{S}| + 2$ and $q = (\delta_{\min})^2(\delta_{\min}/|\mathcal{A}|)^{|\mathcal{S}|}$. Hence, by Lemma 10, for every $\varepsilon > 0$, we can consider $T = n \left\lceil \frac{\log(\varepsilon)}{\log(1-q)} \right\rceil = O\left(|\mathcal{S}||\mathcal{A}|^{|\mathcal{S}|} \log(1/\varepsilon)/\delta_{\min}^{|\mathcal{S}|+2}\right)$, which is at most exponentially large in the input size.

Previous Approaches to Finite Horizon Finite horizon objectives have been studied by a long time. A naive approach to the quantitative analysis for belief-reachability objectives would take the time bound T in Lemma 10 and solve a (fully observable) MDP with $|\mathcal{A} \times \mathcal{S}|^T$ states. This approach implies a 2EXPTIME complexity upper bound. A fundamental result states that, for horizons that are polynomially large with respect to the input, computing the finite horizon value of POMDPs is PSPACE-complete [33, Theorem 6, page 448]. The technique, instead of listing explicitly all exponentially many histories, compactly represents them using nondeterminism and space proportional to the horizon. The conclusion follows from PSPACE being closed under nondeterminism by Savitch's theorem. For exponentially large horizons, this technique leads to an EXPSPACE upper bound. Instead, we prove an EXPTIME upper bound.

Point-based algorithms were introduced for POMDPs by [35] as an alternative to approximating the value of POMDPs by considering a fixed subset of beliefs and projected belief updates. For a survey on point-based algorithms see [39, 42], which focuses on finite horizon objectives but does not provide the complexity upper bound we require.

Finite-Horizon Belief-Reachability Let $T \in \mathbb{N}$ be a finite horizon. The T-step belief-reachability value to $\mathcal{X} \subseteq \mathcal{S}$ is defined by $\mathrm{val}^{T}_{\mathrm{BR}(\mathcal{X})}(b) \coloneqq \max_{\sigma \in \Sigma} \mathbb{P}^{\sigma}_{b}(\exists t \leq T \;\; B_t \in \mathcal{D}_{\mathcal{X}})$.

Lemma 11. *Approximating the T-step belief-reachability value of revealing POMDPs up to an additive error of $\varepsilon > 0$ can be computed in exponential time, formally, in time $O\left(T^{|\mathcal{S}|}|\mathcal{S}|^{|\mathcal{S}|+1}|\mathcal{A}||\mathcal{Z}|/\varepsilon^{(|\mathcal{S}|-1)}\right)$.*

Proof Sketch. We approximate the T-step belief-reachability value by discretizing the belief simplex and applying projected Bellman updates over this grid. Define a k-uniform grid $G_k \subseteq \Delta(\mathcal{S})$ so that every belief $b \in \Delta(\mathcal{S})$ is at L_1-distance at most $\varepsilon/(T+1)$ from some grid point $\Pi_k(b)$, where $k = \lceil (T+1)|\mathcal{S}|/\varepsilon \rceil$. Then, compute the Bellman updates on grid points, carefully projecting back to G_k after each update, for T steps. By induction on the horizon, the error at each step accumulates by at most $\varepsilon/(T+1)$, yielding a total error at most ε. The grid has size $O((T|\mathcal{S}|/\varepsilon)^{|\mathcal{S}|-1})$, and each Bellman update takes $O(|\mathcal{S}|^2|\mathcal{A}||\mathcal{Z}|)$ time, so the total running time is $O\left(T^{|\mathcal{S}|}|\mathcal{S}|^{|\mathcal{S}|+1}|\mathcal{A}||\mathcal{Z}|\varepsilon^{-(|\mathcal{S}|-1)}\right)$, which is exponential in the input size. □

The following result follows from Lemmas 10 and 11.

Theorem 1. *Quantitative analysis for belief-reachability objectives for revealing POMDPs is in* EXPTIME.

5 Parity Objectives

5.1 Reduction to Belief-Reachability Objectives

In the sequel, we show that, for revealing POMDPs, the parity objective value coincides with the belief-reachability to the set of Dirac on the almost-sure winning parity states. This proof uses the standard notion of *end components* in MDPs, introduced in [1]; see also [6, Section 10.6.3].

Underlying MDPs of POMDPs For a POMDP $P = (\mathcal{S}, \mathcal{A}, \mathcal{Z}, \delta, b_0)$, we define its underlying MDP as $M = (\mathcal{S} \times (\mathcal{Z} \cup \{\square\}), \mathcal{A}, \widetilde{\delta}, \widetilde{b}_0)$, where the transition function is defined as, for all $s, s' \in \mathcal{S}, z \in \mathcal{Z} \cup \{\square\}, z' \in \mathcal{Z}$,

$$\widetilde{\delta}\big((s,z),a\big)\big((s',z')\big) \coloneqq \delta(s,a)(s',z')\,,$$

and the initial belief is defined as $\widetilde{b}_0\big((s,\square)\big) \coloneqq b_0(s)$ for all states $s \in \mathcal{S}$. This construction captures the entire dynamic of P. We use this notion to analyze the policies derived from the original POMDP.

End Components Let $M = (\mathcal{S}, \mathcal{A}, \delta, b_0)$ be an MDP. An end component is a pair $\mathcal{U} = (\mathcal{Q}, \mathcal{E})$ where $\mathcal{Q} \subseteq \mathcal{S}$ is a subset of states and $\mathcal{E} \colon \mathcal{Q} \rightrightarrows \mathcal{A}$ assigns each state to a set of nonempty actions such that

- *Closedness.* For all states $q \in \mathcal{Q}$ and actions $a \in \mathcal{E}(q)$, we have $\mathrm{supp}(\delta(q,a)) \subseteq \mathcal{Q}$; and
- *Strong connectivity.* The directed graph with states $\mathcal{Q}$ and edges (q, q') where there exists $a \in \mathcal{E}(q)$ such that $\delta(q,a)(q') > 0$ is strongly connected.

A play ρ visits infinitely often an end component $\mathcal{U} = (\mathcal{Q}, \mathcal{E})$, denoted by $\mathcal{I}(\rho) = (\mathcal{Q}, \mathcal{E})$, if

- The set of states visited infinitely often in ρ is $\mathcal{Q}$; and
- For all state $q \in \mathcal{Q}$, the set of actions played infinitely often when in q is $\mathcal{E}(q)$.

We say $\mathcal{I}(\rho) \subseteq (\mathcal{Q}, \mathcal{E})$ if $\mathcal{I}(\rho) = (\widetilde{\mathcal{Q}}, \widetilde{\mathcal{E}})$, $\widetilde{\mathcal{Q}} \subseteq \mathcal{Q}$ and $\widetilde{\mathcal{E}} \subseteq \mathcal{E}$.

The following statement is a fundamental result of end components in MDPs.

Lemma 12 ([1, Theorems 3.1 and 3.2])**.** *For MDPs, the following assertions hold.*

- *For every end component $\mathcal{U} = (\mathcal{Q}, \mathcal{E})$ and state $q \in \mathcal{Q}$, there exists a policy σ such that*

$$\mathbb{P}^{\sigma}_{\mathbb{1}[q]}\left(\mathcal{I}(\rho) = \mathcal{U}\right) = 1\,.$$

- *For all policies σ, we have*

$$\mathbb{P}^{\sigma}_{b_0}\left(\mathcal{I}(\rho) \text{ is an end component}\right) = 1\,.$$

The following result relates end components in MDPs to policies in general POMDPs.

Lemma 13. *Consider a POMDP P and its underlying MDP M. For every reachable end component $\mathcal{U} = (\mathcal{Q}, \mathcal{E})$ of M, i.e., there exists a policy σ on P such that $\mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = \mathcal{U}) > 0$, we have that, for all states $q \in \mathcal{Q}$, actions $a \in \mathcal{E}(q)$, and signals $z \in \mathcal{Z}$, there exists a policy $\sigma_{q,a,z}$ on P such that*

$$\mathbb{P}^{\sigma_{q,a,z}}_{b}(\mathcal{I}(\rho) \subseteq \mathcal{U}) \;=\; 1\,,$$

where b is the belief after starting with belief $\mathbb{1}[q]$, playing action a, and receiving signal z.

Proof Sketch. Consider a reachable end component $\mathcal{U} = (\mathcal{Q}, \mathcal{E})$ of M with corresponding policy σ, i.e., $\mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = \mathcal{U}) > 0$. On the event $\{\mathcal{I}(\rho) = \mathcal{U}\}$, every state-action pair (q, a) with $q \in \mathcal{Q}$ and $a \in \mathcal{E}(q)$ is taken infinitely often, and each such occurrence results in at least one signal z that occurs with positive probability. Fix $q \in \mathcal{Q}$, $a \in \mathcal{E}(q)$ and a signal z that can follow (q, a). By contradiction, assume that from the belief b obtained after starting with $\mathbb{1}[q]$, playing action a, and receiving signal z, no policy can stay inside $\mathcal{U}$ almost surely. Then, starting from b, any policy has a positive, state–independent strictly positive lower bound on the probability of leaving $\mathcal{Q}$ within a bounded number of steps. Because (q, a) is visited infinitely often on the event $\{\mathcal{I}(\rho) = \mathcal{U}\}$, these lower-bounded leaving probabilities accumulate, driving the probability of ever leaving $\mathcal{Q}$ to 1. This contradicts $\mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = \mathcal{U}) > 0$. Hence, such a leaving bound cannot exist: there must be a policy $\sigma_{q,a,z}$ that, from b, keeps the play inside $\mathcal{Q}$ (and therefore inside $\mathcal{U}$) with probability 1. □

The following example demonstrates that the above result is tight in the sense that it can not guarantee a policy $\sigma_{q,a,z}$ such that $\mathbb{P}^{\sigma_{q,a,z}}_{b}\big(\mathcal{I}(\rho) = \mathcal{U}\big) = 1$.

Example 1. *Consider the example presented in [18, Example 4.3] of a POMDP with only one possible signal in Figure 1, commonly known as a blind MDP. It has four states $\mathcal{S} = \{s_0, s_1, \bot, \top\}$ and two actions:* wait *(w) and* commit *(c), i.e., $\mathcal{A} = \{w, c\}$. The priority function is defined as*

$$\mathrm{pri}(s_0) \coloneqq 1,\ \mathrm{pri}(s_1) \coloneqq 1,\ \mathrm{pri}(\bot) \coloneqq 1,\ \mathrm{pri}(\top) = 0\,.$$

The transitions are defined as follows. (a) Under action w, state s_0 moves either to s_1 or loops, both with probability $1/2$; states s_1 and $\bot$ loop; and state $\top$ moves to s_0. (b) Under action c, state s_0 moves to $\bot$; state s_1 moves $\top$; state $\bot$ loops; and state $\top$ moves to s_0. The only end component of the underlying MDP that satisfies the parity condition is

$$\mathcal{U} = \big(\{s_1, \top\},\ \{(s_0, w), (s_1, w), (s_1, c), (\top, w), (\top, c)\}\big)\,.$$

Note that, for every ε-optimal policy σ, we have that $\mathbb{P}^{\sigma}_{\mathbb{1}[s_0]}(\mathcal{I}(\rho) = \mathcal{U}) \geq 1 - \varepsilon$. This is achieved only by policies that wait long enough before committing, which requires unbounded

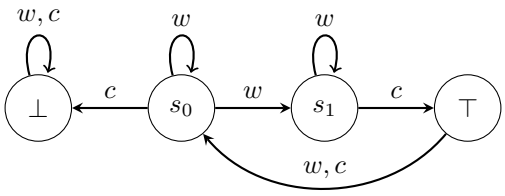

Figure 1: Example of a classic general POMDP that requires infinite memory policies for parity objectives. Edges represent a positive probability transition between states when the corresponding action in its label is used.

memory. Conversely, no policy can guarantee $\mathbb{P}^{\sigma}_{\mathbb{1}[s_0]}(\mathcal{I}(\rho) = \mathcal{U}) = 1$. *Indeed, whenever action c is taken, the belief on state* s_0 *is positive and therefore the state absorbs at* $\bot$ *with positive probability. The best that can be achieved almost surely is to stay forever in the end component* $\widetilde{\mathcal{U}} = \big(\{s_1\}, \{(s_1, w)\}\big)$, *whose minimal priority is 1 and therefore violates the parity condition.* □

We now show that in revealing POMDPs, the above result can be extended to guarantee a policy σ_q such that $\mathbb{P}^{\sigma_q}_{\mathbb{1}[q]}\big(\mathcal{I}(\rho) = \mathcal{U}\big) = 1$.

Lemma 14. *Consider a revealing POMDP P and its underlying MDP M. For every policy* σ *on P and end component* $\mathcal{U} = (\mathcal{Q}, \mathcal{E})$ *of M such that* $\mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = \mathcal{U}) > 0$, *we have, for all states* $q \in \mathcal{Q}$, *there exists a policy* σ_q *on P such that*

$$\mathbb{P}^{\sigma_q}_{\mathbb{1}[q]}(\mathcal{I}(\rho) = \mathcal{U}) \;=\; 1\,.$$

Proof Sketch. Consider a policy σ that $\mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = \mathcal{U}) > 0$. Fix a state $q \in \mathcal{Q}$. We construct a policy σ_q that proceeds in *phases*, one for each ordered pair $(\widetilde{q}, \widetilde{a})$ where $\widetilde{q} \in \mathcal{Q}$ and $\widetilde{a} \in \mathcal{E}(\widetilde{q})$. The goal of a phase is to reach $\widetilde{q}$ and to play action $\widetilde{a}$. When the current belief is non–Dirac, the policy follows the almost-sure safety policy from Lemma 13 that stays inside $\mathcal{U}$ until a Dirac belief $\mathbb{1}[q']$ is reached, which is guaranteed in finite time because P is revealing. From q', the policy moves inside $\mathcal{U}$ along a fixed finite path of "safe" actions to the target state $\widetilde{q}$; if the belief becomes non-Dirac, the policy returns to the safety policy and retries. Once the belief is $\mathbb{1}[\widetilde{q}]$, the policy plays the action $\widetilde{a}$; upon observing the resulting signal z, switches to the safety policy and starts the next phase. Because every ordered pair $(q, a) \in \mathcal{E}$ is visited infinitely often, each edge of $\mathcal{U}$ is taken infinitely often. The safety policies ensure the play never leaves $\mathcal{Q}$. Consequently, the event $\{\mathcal{I}(\rho) = \mathcal{U}\}$ is satisfied with probability 1 under σ_q when starting from the belief $\mathbb{1}[q]$. □

We are ready to present the reduction of parity to belief-reachability of the almost-sure winning states for parity.

Lemma 15. *Consider a revealing POMDP with parity objectives. Denote the set of states for which, if the initial belief were a Dirac on that state, then the parity condition can be satisfied almost-surely by* $\mathcal{X} \coloneqq \Big\{s \in \mathcal{S} : \exists \sigma \in \Sigma \quad \mathbb{P}^{\sigma}_{\mathbb{1}[s]}(\text{Parity}) = 1\Big\}$. *Then, the parity value coincides with the belief-reachability value to* $\mathcal{X}$, *i.e., for all beliefs* $b \in \Delta(\mathcal{S})$,

$$\text{val}_{\text{P}}(b) = \text{val}_{\text{BR}(\mathcal{X})}(b)\,.$$

Proof sketch. Consider a policy $\sigma \in \Sigma$. Consider its end components in the underlying MDP on $\mathcal{S} \times \mathcal{Z}$. Note that each end component either does satisfy or does not satisfy the parity condition. Moreover, if they satisfy the parity condition, then σ is an almost-sure winning policy

starting from any state inside the end component. Therefore, the probability of satisfying the parity condition under σ corresponds to the belief-reachability to the end components where the parity condition is satisfied. In particular, in these end components, parity condition is satisfied almost-surely. We conclude because the policy is arbitrary. □

5.2 Proofs of Main Results

Building on Lemma 15, which reduces parity to belief-reachability, and the EXPTIME procedure in Theorem 1 for approximating the belief-reachability value, we now (a) obtain an EXPTIME algorithm for approximating parity values (Theorem 3); and (b) show that limit-sure and almost-sure winning coincide (Theorem 4).

Theorem 3. *Quantitative analysis for parity objectives for revealing POMDPs is in* EXPTIME.

Proof Sketch. By [7, Theorem 3], computing almost-sure winning parity states in the revealing POMDP is EXPTIME. Therefore, this result is a direct implication of Lemma 15 and Theorem 1. □

Theorem 4. *Optimal policies exist for parity objectives for revealing POMDPs.*

Proof Sketch. It is a direct implication of Corollary 9 and Lemma 15. □

Concluding Remarks In this work, we consider revealing POMDPs which have been studied in the literature and provide decidability results for the fundamental computational problems for this model. Interesting directions for future work include exploring the practical applicability of our algorithms and extending our decidability results to other classes of POMDPs.

Acknowledgements

This work was supported by the ANRT under the French CIFRE Ph.D program in collaboration between NyxAir and Paris-Dauphine University (Contract: CIFRE N° 2022/0513), by the French Agence Nationale de la Recherche (ANR) under reference ANR-21-CE40-0020 (CONVERGENCE project), by Austrian Science Fund (FWF) 10.55776/COE12, and by the ERC CoG 863818 (ForM-SMArt) grant.

References

[1] Alfaro, L. 1998. *Formal Verification of Probabilistic Systems*. Ph.D. diss., Stanford University, Stanford, CA, USA.

[2] Andersson, D.; and Miltersen, P. B. 2009. The Complexity of Solving Stochastic Games on Graphs. In *Algorithms and Computation*, volume 5878, 112–121. Berlin, Heidelberg: Springer.

[3] Avrachenkov, K.; Dhiman, M.; and Kavitha, V. 2025. Constrained Average-Reward Intermittently Observable MDPs. arXiv:2504.13823.

[4] Baier, C.; Bertrand, N.; and Größer, M. 2008. On Decision Problems for Probabilistic Büchi Automata. In *Foundations of Software Science and Computational Structures*, volume 4962, 287–301. Berlin, Heidelberg: Springer.

[5] Baier, C.; Größer, M.; and Bertrand, N. 2012. Probabilistic ω-automata. *Journal of the ACM (JACM)*, 59(1): 1–52.

[6] Baier, C.; and Katoen, J.-P. 2008. *Principles of Model Checking*. Cambridge, MA, USA: MIT press.

[7] Belly, M.; Fijalkow, N.; Gimbert, H.; Horn, F.; Pérez, G. A.; and Vandenhove, P. 2025. Revelations: A Decidable Class of POMDPs with Omega-Regular Objectives. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25): 26454–26462.

[8] Bertsekas, D. P. 1976. *Dynamic Programming and Stochastic Control*. New York: Academic Press.

[9] Billingsley, P. 2012. *Probability and Measure*. Hoboken, NJ, USA: Wiley.

[10] Bonet, B.; and Geffner, H. 2009. Solving POMDPs: RTDP-Bel vs. Point-based Algorithms. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI'09, 1641–1646. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[11] Chatterjee, K.; Chmelik, M.; Gupta, R.; and Kanodia, A. 2015. Qualitative Analysis of POMDPs with Temporal Logic Specifications for Robotics Applications. In *IEEE International Conference on Robotics and Automation (ICRA)*, 325–330. Seattle, WA, USA: IEEE.

[12] Chatterjee, K.; Chmelík, 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.

[13] Chatterjee, K.; Doyen, L.; Gimbert, H.; and Henzinger, T. A. 2010. Randomness for Free. In *Proceedings of the 35th International Conference on Mathematical Foundations of Computer Science*, 246–257. Berlin, Heidelberg: Springer.

[14] Chatterjee, K.; Doyen, L.; and Henzinger, T. A. 2010. Qualitative Analysis of Partially-Observable Markov Decision Processes. In *International Symposium on Mathematical Foundations of Computer Science*, 258–269. Berlin, Heidelberg: Springer.

[15] Chatterjee, K.; Doyen, L.; Raskin, J.-F.; and Sankur, O. 2025. The Value Problem for Multiple-Environment MDPs with Parity Objective. In *52nd International Colloquium on Automata, Languages, and Programming (ICALP 2025)*, volume 334 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 150:1–150:17. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[16] Chatterjee, K.; and Henzinger, T. A. 2010. Probabilistic Automata on Infinite Words: Decidability and Undecidability Results. In *International Symposium on Automated Technology for Verification and Analysis*, 1–16. Berlin, Heidelberg: Springer.

[17] Chatterjee, K.; Lurie, D.; Saona, R.; and Ziliotto, B. 2024. Ergodic Unobservable MDPs: Decidability of Approximation. arXiv:2405.12583.

[18] Chatterjee, K.; Saona, R.; and Ziliotto, B. 2021. Finite-Memory Strategies in POMDPs with Long-Run Average Objectives. *Mathematics of Operations Research*, 47(1): 100–119.

[19] Chatterjee, K.; and Tracol, M. 2012. Decidable Problems for Probabilistic Automata on Infinite Words. In *27th Annual IEEE Symposium on Logic in Computer Science*, 185–194. Dubrovnik, Croatia: IEEE.

[20] Chen, F.; Wang, H.; Xiong, C.; Mei, S.; and Bai, Y. 2023. Lower Bounds for Learning in Revealing POMDPs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, 5104–5161. Honolulu, Hawaii, USA: JMLR.org.

[21] Chen, G.; and Liew, S.-C. 2023. Intermittently Observable Markov Decision Processes. arXiv:2302.11761.

[22] de Alfaro, L.; Faella, M.; Majumdar, R.; and Raman, V. 2005. Code Aware Resource Management. In *Proceedings of the 5th ACM International Conference on Embedded Software*, EMSOFT '05, 191–202. New York, NY, USA: Association for Computing Machinery.

[23] Durbin, R.; Eddy, S. R.; Krogh, A.; and Mitchison, G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press.

[24] Fijalkow, N.; Gimbert, H.; Kelmendi, E.; and Oualhadj, Y. 2015. Deciding the Value 1 Problem for Probabilistic Leaktight Automata. *Logical Methods in Computer Science*, 11.

[25] Filar, J.; and Vrieze, K. 1997. *Competitive Markov Decision Processes*. New York, NY, USA: Springer.

[26] Gimbert, H.; and Oualhadj, Y. 2010. Probabilistic Automata on Finite Words: Decidable and Undecidable Problems. In *Automata, Languages and Programming*, volume 6199, 527–538. Berlin, Heidelberg: Springer.

[27] Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and Acting in Partially Observable Stochastic Domains. *Artificial Intelligence*, 101(1-2): 99–134.

[28] Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4: 237–285.

[29] Kechris, A. S. 1995. *Classical Descriptive Set Theory*. New York, NY, USA: Springer.

[30] Kress-Gazit, H.; Fainekos, G. E.; and Pappas, G. J. 2009. Temporal-Logic-Based Reactive Mission and Motion Planning. *IEEE Transactions on Robotics*, 25(6): 1370–1381.

[31] Liu, Q.; Chung, A.; Szepesvari, C.; and Jin, C. 2022. When Is Partially Observable Reinforcement Learning Not Scary? In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, 5175–5220. PMLR.

[32] 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.

[33] Papadimitriou, C. H.; and Tsitsiklis, J. N. 1987. The Complexity of Markov Decision Processes. *Mathematics of Operations Research*, 12(3): 441–450.

[34] Paz, A. 1971. *Introduction to Probabilistic Automata*. New York, NY, USA: Academic Press.

[35] Pineau, J.; Gordon, G.; and Thrun, S. 2003. Point-Based Value Iteration: An Anytime Algorithm for POMDPs. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 1025–1030. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[36] Pogosyants, A.; Segala, R.; and Lynch, N. 2000. Verification of the Randomized Consensus Algorithm of Aspnes and Herlihy: a Case Study. *Distributed Computing*, 13(3): 155–186.

[37] Puterman, M. L. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley.

[38] Rabin, M. O. 1963. Probabilistic Automata. *Information and control*, 6(3): 230–245.

[39] Shani, G.; Pineau, J.; and Kaplow, R. 2013. A Survey of Point-Based POMDP Solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1): 1–51.

[40] Thomas, W. 1997. Languages, Automata, and Logic. In *Handbook of Formal Languages: Volume 3 Beyond Words*, 389–455. Berlin, Heidelberg: Springer.

[41] Van Der Vegt, M.; Jansen, N.; and Junges, S. 2023. Robust Almost-Sure Reachability in Multi-Environment MDPs. In *Tools and Algorithms for the Construction and Analysis of Systems*, volume 13993, 508–526. Berlin, Heidelberg: Springer.

[42] Walraven, E.; and Spaan, M. T. J. 2019. Point-Based Value Iteration for Finite-Horizon POMDPs. *Journal of Artificial Intelligence Research*, 65: 307–341.

A Proofs of Section 4

Proposition 6. *The set of reliable actions is nonempty.*

Proof. Consider a POMDP with initial belief $b_0 \in \Delta(\mathcal{S})$ and set of target states $\mathcal{X} \subseteq \mathcal{S}$. By definition of the value, we have that $\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(b_0) = \max_{a \in \mathcal{A}} \mathbb{E}^a_{b_0}(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_1))$. Because the set of actions $\mathcal{A}$ is finite, there exists $a \in \mathcal{A}$ that achieves the maximum, which is a reliable action. □

Lemma 7. *Consider a POMDP with belief-reachability objectives. If a policy is stopping and uses only reliable actions, then it is optimal.*

Proof. Consider a POMDP with initial belief $b_0 \in \Delta(\mathcal{S})$ and belief-reachability objectives. Denote the set of target states and the corresponding set of terminal states by $\mathcal{X} \subseteq \mathcal{S}$ and $\mathcal{T} \subseteq \mathcal{S}$, respectively. Consider a policy $\sigma \in \Sigma$ that is stopping and uses only reliable actions. We show that σ is optimal.

Denote the hitting time of terminal Dirac beliefs by $\tau \coloneqq \inf\{t \geq 0 : B_t \in \mathcal{D}_{\mathcal{T}}\}$. Because the policy σ is (n, q)-stopping, for every belief $b \in \Delta(\mathcal{S})$, we have that $\mathbb{P}^\sigma_b(\tau \leq n) \geq q$. Therefore, for every $k \geq 1$,

$$\mathbb{P}^\sigma_b(\tau \leq kn) \geq 1 - (1-q)^k \,.$$

Taking the limit as k goes to infinity, we deduce that $\mathbb{P}^\sigma_b(\tau < \infty) = 1$.

Because σ uses only reliable actions, we have that, for every $t \geq 1$,

$$\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(b) = \mathbb{E}^\sigma_b\left(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_t)\right) .$$

Therefore, for every $T \geq 1$ and $b \in \Delta(\mathcal{S})$,

$$\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(b) = \mathbb{E}^\sigma_b\left(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_\tau)\mathbb{1}_{\{\tau \leq T\}}\right) + \mathbb{E}^\sigma_b\left(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_\tau)\mathbb{1}_{\{\tau > T\}}\right) .$$

Whenever $\tau \leq T$, we have that there exists $t \leq T$ such that $B_t \in \mathcal{D}_{\mathcal{T}}$, and thus $\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_t) = \mathbb{1}_{\{B_t \in \mathcal{D}_{\mathcal{X}}\}}$. Taking T to infinity,

$$\begin{aligned}
\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(b) &= \lim_{T\to\infty} \mathbb{E}^\sigma_b\left(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_\tau)\mathbb{1}_{\{\tau \leq T\}}\right) \\
&= \lim_{T\to\infty} \mathbb{P}^\sigma_b(\exists t \leq T \quad B_t \in \mathcal{D}_{\mathcal{X}}) \\
&= \mathbb{P}^\sigma_b(\exists t \geq 0 \quad B_t \in \mathcal{D}_{\mathcal{X}}) \,.
\end{aligned}$$

Therefore, the policy σ is optimal. □

Lemma 8. *Consider a revealing POMDP with belief-reachability objectives. The policy σ that plays reliable actions uniformly at random is (n, q)-stopping with parameters $n \coloneqq |\mathcal{S}| + 2$ and $q \coloneqq \delta^2_{\min}\left(\delta_{\min}/|\mathcal{A}|\right)^{|\mathcal{S}|}$.*

Proof. Consider a revealing POMDP P with belief-reachability objectives. Recall that, by Proposition 6, every belief has at least one reliable action, i.e., for every $b \in \Delta(\mathcal{S})$, we have that $\mathcal{R}(b) \neq \emptyset$. Let σ be the policy that plays uniformly at random among all reliable actions. We show that σ is (n, q)-stopping, i.e., for every initial belief $b \in \Delta(\mathcal{S})$,

$$\mathbb{P}^\sigma_b(\exists t \leq n \quad B_t \in \mathcal{D}_{\mathcal{T}}) \geq q \,,$$

where

$$n = |\mathcal{S}| + 2 \quad \text{and} \quad q = \delta^2_{\min}\left(\frac{\delta_{\min}}{|\mathcal{A}|}\right)^{|\mathcal{S}|} .$$

Denote the set of terminal states by $\mathcal{T} \subseteq \mathcal{S}$. Define the set of states that can reach $\mathcal{T}$ using only reliable actions, indexed by the number of steps required, as follows. First, $\mathcal{S}_0 \coloneqq \mathcal{T}$ and, for $t \geq 1$,

$$\mathcal{S}_t \coloneqq \left\{ s \in \mathcal{S} : \exists a \in \mathcal{R}\left(\mathbb{1}[s]\right) \quad \sum_{\widetilde{s} \in \mathcal{S}_{t-1}, z \in \mathcal{Z}} \delta(s,a)\left(\widetilde{s}, z\right) > 0 \right\}.$$

We claim that at most $|\mathcal{S}|$ steps are required to cover all states, i.e., $\mathcal{S}_{|\mathcal{S}|} = \mathcal{S}$.

Indeed, by contradiction, assume that the set of uncovered states $\mathcal{L} \coloneqq \mathcal{S} \setminus \mathcal{S}_{|\mathcal{S}|}$ is nonempty. Note that $(\mathcal{S}_t)_{t \geq 0}$ is an increasing and bounded sequence and therefore reaches a fixpoint equal to $\mathcal{S}_{|\mathcal{S}|}$. On the one hand, for all state $s \in \mathcal{L}$, under reliable actions for the belief $\mathbb{1}[s]$, the state can not move from $\mathcal{L}$ to $\mathcal{S}_{|\mathcal{S}|}$. Formally, for all $s \in \mathcal{L}, a \in \mathcal{R}(\mathbb{1}[s])$,

$$\operatorname{supp}(\delta(s,a)) \subseteq \mathcal{L} \times \mathcal{Z}\,.$$

On the other hand, using unreliable actions imply decreasing the future value, i.e., for all $s \in \mathcal{L}$,

$$\max_{a \in \mathcal{A} \setminus \mathcal{R}(\mathbb{1}[s])} \mathbb{E}^a_{\mathbb{1}[s]}(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_1)) < \mathrm{val}_{\mathrm{BR}(\mathcal{X})}(\mathbb{1}[s])\,.$$

We show that there is a state in $\mathcal{L}$ with no ε-optimal policy for ε small enough, which is a contradiction.

Consider the state $s \in \mathcal{L}$ with the highest belief-reachability value. Note that states with value zero are terminal, so s has strictly positive value. Denote the minimum gap obtained by using unreliable actions by

$$\zeta \coloneqq \min_{\substack{s \in \mathcal{L}, \\ a \notin \mathcal{R}(\mathbb{1}[s])}} \mathrm{val}_{\mathrm{BR}(\mathcal{X})}(\mathbb{1}[s]) - \mathbb{E}^a_{\mathbb{1}[s]}(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_1)) > 0\,.$$

Take $\varepsilon \in (0, \min\{\zeta, \mathrm{val}_{\mathrm{BR}(\mathcal{X})}(\mathbb{1}[s])\})$ and consider a policy σ_ε that is ε-optimal policy for the initial belief $\mathbb{1}[s]$. Note that, because reliable actions do not connect $\mathcal{L}$ to $\mathcal{T}$ and $\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(\mathbb{1}[s]) > 0$, the policy σ_ε must necessarily use unreliable actions. We show that σ_ε is not ε-optimal, leading to a contradiction.

Denote the first time the action played has positive probability of leaving the set $\mathcal{L}$ by

$$\tau \coloneqq \inf\left\{ t \geq 0 \;:\; \operatorname{supp}\big(\delta(S_t, A_t)\big) \notin \mathcal{L} \times \mathcal{Z} \right\}.$$

Then, we bound the belief-reachability guaranteed by σ_ε as follows.

$$\begin{aligned}
\mathbb{P}^{\sigma_\varepsilon}_{\mathbb{1}[s]}(\exists t \in \mathbb{N} \quad B_t \in \mathcal{D}_{\mathcal{X}}) &= \mathbb{E}^{\sigma_\varepsilon}_{\mathbb{1}[s]}\left(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(B_{\tau+1}) \mathbb{1}[\tau < \infty]\right) && (\mathcal{X} \cap \mathcal{L} = \emptyset) \\
&\leq \mathbb{E}^{\sigma_\varepsilon}_{\mathbb{1}[s]}\left(\mathrm{val}_{\mathrm{BR}(\mathcal{X})}\left(\mathbb{1}[S_\tau]\right) \mathbb{1}[\tau < \infty]\right) - \zeta && (\text{Definition of } \zeta) \\
&\leq \mathrm{val}_{\mathrm{BR}(\mathcal{X})}\left(\mathbb{1}[s]\right) - \zeta && (\text{Choice of } s) \\
&< \mathrm{val}_{\mathrm{BR}(\mathcal{X})}(\mathbb{1}[s]) - \varepsilon\,, && (\varepsilon < \zeta)
\end{aligned}$$

which contradicts the fact that σ_ε is ε-optimal. We conclude that $\mathcal{S}_{|\mathcal{S}|} = \mathcal{S}$.

We now prove that the policy σ is stopping with parameters $n = |\mathcal{S}|+2$ and $q = \delta^2_{\min}\left(\frac{\delta_{\min}}{|\mathcal{A}|}\right)^{|\mathcal{S}|}$. Note that, because P is revealing, for all $t \in [|\mathcal{S}|]$,

$$s \in \mathcal{S}_t \quad \Rightarrow \quad \exists s' \in \mathcal{S}_{t-1}, a \in \mathcal{R}(\mathbb{1}[s]) \quad \delta(s,a)(s',s') > 0\,.$$

By recursion and because σ plays all reliable actions uniformly at random, we conclude that, for all $s \in \mathcal{S}$,

$$\mathbb{P}^{\sigma}_{\mathbb{1}_{[s]}}\left(\exists t \leq |\mathcal{S}| \quad S_t \in \mathcal{T}\right) \geq \left(\frac{\delta_{\min}}{|\mathcal{A}|}\right)^{|\mathcal{S}|}.$$

and thus,

$$\mathbb{P}^{\sigma}_{\mathbb{1}_{[s]}}\left(\exists t \leq |\mathcal{S}|+1 \quad B_t \in \mathcal{D}_{\mathcal{T}}\right) \geq \delta_{\min}\left(\frac{\delta_{\min}}{|\mathcal{A}|}\right)^{|\mathcal{S}|}.$$

To conclude, for all initial beliefs $b \in \Delta(\mathcal{S})$, after every action, the probability that the next belief is a Dirac at least $\delta_{\min}$. □

Lemma 10. *Consider a POMDP with belief-reachability objectives and an (n,q)-stopping optimal policy σ. For all $\varepsilon > 0$, the finite horizon $T \coloneqq n\left\lceil \frac{\log(\varepsilon)}{\log(1-q)} \right\rceil$ is such that*

$$\mathbb{P}^{\sigma}_{b_0}(\exists t \leq T \quad B_t \in \mathcal{D}_{\mathcal{T}}) \geq \mathrm{val}_{\mathrm{BR}}(b_0) - \varepsilon\,.$$

Proof. Consider a POMDP with belief-reachability objectives and an (n,q)-stopping optimal policy σ. Denote the hitting time of the terminal set $\mathcal{T} \subseteq \mathcal{S}$ by

$$\tau \coloneqq \inf\{t \geq 0 : B_t \in \mathcal{D}_{\mathcal{T}}\}.$$

Because σ is (n,q)-stopping, we have that $\mathbb{P}^{\sigma}_{b_0}(\tau \leq n) \geq q$. Similarly, considering blocks of size n, for every integer $k \geq 1$, we have that $\mathbb{P}^{\sigma}_{b_0}(\tau > kn) \leq (1-q)^k$. Thus, τ is finite $\mathbb{P}^{\sigma}_{b_0}$-a.s.

For every $b_0 \in \Delta(\mathcal{S})$, we have that

$$\begin{aligned}
\mathrm{val}_{\mathrm{BR}(\mathcal{X})}(b_0) &= \mathbb{P}^{\sigma}_{b_0}(\exists t \geq 0 \quad B_t \in \mathcal{D}_{\mathcal{X}}) && \left(\text{Definition of } \mathrm{val}_{\mathrm{BR}(\mathcal{X})}\right)\\
&= \mathbb{P}^{\sigma}_{b_0}(B_\tau \in \mathcal{D}_{\mathcal{X}}) && (\tau < \infty)\\
&= \mathbb{P}^{\sigma}_{b_0}(\tau \leq T, B_\tau \in \mathcal{D}_{\mathcal{X}}) + \mathbb{P}^{\sigma}_{b_0}(\tau > T, B_\tau \in \mathcal{D}_{\mathcal{X}}) && (\text{Partitioning})\\
&\leq \mathbb{P}^{\sigma}_{b_0}(\tau \leq T, B_\tau \in \mathcal{D}_{\mathcal{X}}) + \mathbb{P}^{\sigma}_{b_0}(\tau > T) && (\text{Inclusion})\\
&\leq \mathbb{P}^{\sigma}_{b_0}(\exists t \leq T \quad B_t \in \mathcal{D}_{\mathcal{X}}) + \mathbb{P}^{\sigma}_{b_0}(\tau > T)\,. && (\text{Definition of } \tau)
\end{aligned}$$

We deduce that $\mathbb{P}^{\sigma}_{b_0}(\tau > T) \leq \varepsilon$ by taking $T = n\left\lceil \frac{\log(\varepsilon)}{\log(1-q)} \right\rceil$, which concludes the proof. □

Lemma 11. *Approximating the T-step belief-reachability value of revealing POMDPs up to an additive error of $\varepsilon > 0$ can be computed in exponential time, formally, in time $O\left(T^{|\mathcal{S}|}|\mathcal{S}|^{|\mathcal{S}|+1}|\mathcal{A}||\mathcal{Z}|/\varepsilon^{(|\mathcal{S}|-1)}\right)$.*

Proof. Consider a revealing POMDP with a set of target states $\mathcal{X} \subseteq \mathcal{S}$ and a finite horizon $T \in \mathbb{N}$. Denote the corresponding set of terminal states by $\mathcal{T} \subseteq \mathcal{S}$. Consider $b \in \Delta(\mathcal{S})$ and denote the L1-norm by $\|b\|_1 \coloneqq \sum_{s\in\mathcal{S}} |b(s)|$.

Simplex Discretization: We first discretize the set of beliefs $\Delta(\mathcal{S})$. Let $k \geq 1$ and $\varepsilon > 0$. Define the k-uniform grid of beliefs

$$G_k \coloneqq \left\{b \in \Delta(\mathcal{S}) : b(s) = \frac{n(s)}{k}, n(s) \in \{0,\dots,k\}, \sum_{s=1}^{|\mathcal{S}|} n(s) = k\right\}.$$

Denote $\Pi_k \colon \Delta(\mathcal{S}) \to G_k$ the projection function defined by

$$\Pi_k(b) \coloneqq \operatorname*{arg\,min}_{\substack{b' \in G_k \\ \mathrm{supp}(b)=\mathrm{supp}(b')}} \|b - b'\|_1.$$

By construction, it holds that, for every $b \in \Delta(\mathcal{S})$, there exists $\Pi_k(b) \in G_k$ with

$$\|b - \Pi_k(b)\|_1 \leq \frac{|\mathcal{S}|}{k},$$

and taking $k = \left\lceil \frac{(T+1)|\mathcal{S}|}{\varepsilon} \right\rceil$, we have $\|b - \Pi_k(b)\|_1 \leq \frac{\varepsilon}{T+1}$.

Bellman Equations: We now define the Bellman equations for the T-step belief-reachability value and then for the T-step belief-reachability value on the grid previously defined. For every $t \leq T$ and $b \in \Delta(\mathcal{S})$, define the bellman equation of the t-step belief reachability value in P by

- Base case: we have that

$$\mathrm{val}^0_{\mathrm{BR}(\mathcal{X})}(b) \coloneqq \begin{cases} 1 & \text{if } b = \mathbb{1}[s] \text{ for some } s \in \mathcal{X}, \\ 0 & \text{otherwise.} \end{cases}$$

- Induction case: we have that

$$\mathrm{val}^t_{\mathrm{BR}(\mathcal{X})}(b) \coloneqq \mathbb{1}[b \in \mathcal{D}_{\mathcal{X}}] + (1 - \mathbb{1}[b \in \mathcal{D}_{\mathcal{X}}]) \max_{a \in \mathcal{A}} \sum_{z \in \mathcal{Z}} \sum_{s,s' \in \mathcal{S}} b(s)\delta(s,a)(s',z)\, \mathrm{val}^{t-1}_{\mathrm{BR}(\mathcal{X})}(\tau(b,a,z)),$$

  where $\tau(b,a,z)(s') \coloneqq \frac{\sum_{s \in \mathcal{S}} b(s)\delta(s,a)(s',z)}{\sum_{s,s' \in \mathcal{S}} b(s)\delta(s,a)(s',z)}$.

We now define the value iteration that considers only beliefs on the grid G_k. More formally, for every $b \in \Delta(\mathcal{S})$, there exists $\Pi_k(b) \in G_k$ such that,

- Base case: we have that

$$\widetilde{\mathrm{val}}^0_{\mathrm{BR}(\mathcal{X})}(\Pi_k(b)) \coloneqq \begin{cases} 1 & \text{if } \Pi_k(b) = \mathbb{1}[s] \text{ for some } s \in \mathcal{X}, \\ 0 & \text{otherwise.} \end{cases}$$

- Induction case: we have that

$$\widetilde{\mathrm{val}}^t_{\mathrm{BR}(\mathcal{X})}(\Pi_k(b)) \coloneqq \mathbb{1}[\Pi_k(b) \in \mathcal{D}_{\mathcal{X}}] + (1 - \mathbb{1}[\Pi_k(b) \in \mathcal{D}_{\mathcal{X}}]) \max_{a \in \mathcal{A}} \mathbb{E}^a_{\Pi_k(b)} \left(\widetilde{\mathrm{val}}^{t-1}_{\mathrm{BR}(\mathcal{X})}(\Pi_k(B_1)) \right).$$

Lipschitz Property: We show that the T-step belief-reachability value is a Lipschitz function over beliefs with same support. For all beliefs $b, b' \in \Delta(\mathcal{S})$ with $\mathrm{supp}(b) = \mathrm{supp}(b')$, we have

$$\begin{aligned}
&\left| \mathrm{val}^T_{\mathrm{BR}(\mathcal{X})}(b) - \mathrm{val}^T_{\mathrm{BR}(\mathcal{X})}(b') \right| \\
&\quad = \left| \max_{a \in \mathcal{A}} \mathbb{E}^a_b \left(\mathrm{val}^{T-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) - \max_{a \in \mathcal{A}} \mathbb{E}^a_{b'} \left(\mathrm{val}^{T-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) \right| && \text{(Bellman optimality)} \\
&\quad \leq \left| \max_{a \in \mathcal{A}} \left[\mathbb{E}^a_b \left(\mathrm{val}^{T-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) - \mathbb{E}^a_{b'} \left(\mathrm{val}^{T-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) \right] \right| && \text{(Subadditivity of max)} \\
&\quad \leq \max_{a \in \mathcal{A}} \left| \mathbb{E}^a_b \left(\mathrm{val}^{T-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) - \mathbb{E}^a_{b'} \left(\mathrm{val}^{T-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) \right| && \text{(Monotonicity of max)} \\
&\quad = \max_{a \in \mathcal{A}} \left| \sum_{s \in \mathcal{S}} \big(b(s) - b'(s)\big) \mathbb{E}^a_{1[s]} \left(\mathrm{val}^{T-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) \right| && \left(\text{Affinity of } \mathbb{E}^a_b \left(\mathrm{val}^{T-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) \right) \\
&\quad \leq \sum_{s \in \mathcal{S}} |b(s) - b'(s)| = \|b - b'\|_1 , && \left(0 \leq \mathbb{E}^a_{1[s]} \left(\mathrm{val}^{T-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) \leq 1 \right)
\end{aligned}$$

Error Bound: Take $k = \left\lceil \frac{(T+1)|\mathcal{S}|}{\varepsilon} \right\rceil$. Define for every $t \in \{0, \dots, T\}$ and belief $b \in \Delta(\mathcal{S})$,

$$E_t(b) \coloneqq \left| \mathrm{val}^t_{\mathrm{BR}(\mathcal{X})}(b) - \widetilde{\mathrm{val}}^t_{\mathrm{BR}(\mathcal{X})}(\Pi_k(b)) \right|$$

We prove using an induction argument that, for every $t \leq T$ and $b \in \Delta(\mathcal{S})$,

$$E_t(b) \leq \frac{(t+1)\varepsilon}{T+1}.$$

Observe that when $t = 0$, we have that $E_0(b) \leq \frac{\varepsilon}{T+1}$. Now, assume that $E_{t-1} \leq \frac{t\varepsilon}{T+1}$. Therefore, we have that, for every $b \in \Delta(\mathcal{S})$ and taking $k = \left\lceil \frac{(T+1)|\mathcal{S}|}{\varepsilon} \right\rceil$,

$$\begin{aligned}
E_t(b) &= \left| \mathrm{val}^t_{\mathrm{BR}(\mathcal{X})}(b) - \widetilde{\mathrm{val}}^t_{\mathrm{BR}(\mathcal{X})}(\Pi_k(b)) \right| \\
&\leq \max_{a \in \mathcal{A}} \left| \mathbb{E}^a_b \left(\mathrm{val}^{t-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) - \mathbb{E}^a_{\Pi_k(b)} \left(\widetilde{\mathrm{val}}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(\Pi_k(B_1) \right) \right) \right| \\
&\leq \max_{a \in \mathcal{A}} \left| \mathbb{E}^a_b \left(\mathrm{val}^{t-1}_{\mathrm{BR}(\mathcal{X})}(B_1) \right) - \mathbb{E}^a_{\Pi_k(b)} \left(\mathrm{val}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(B_1 \right) \right) \right| \\
&\quad + \max_{a \in \mathcal{A}} \left| \mathbb{E}^a_{\Pi_k(b)} \left(\mathrm{val}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(B_1 \right) \right) - \mathbb{E}^a_{\Pi_k(b)} \left(\widetilde{\mathrm{val}}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(\Pi_k(B_1) \right) \right) \right| \\
&\leq \|b - \Pi_k(b)\|_1 && \text{(Lipschitz property)} \\
&\quad + \max_{a \in \mathcal{A}} \left| \mathbb{E}^a_{\Pi_k(b)} \left(\mathrm{val}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(B_1 \right) \right) - \mathbb{E}^a_{\Pi_k(b)} \left(\widetilde{\mathrm{val}}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(\Pi_k(B_1) \right) \right) \right| \\
&\leq \|b - \Pi_k(b)\|_1 \\
&\quad + \max_{a \in \mathcal{A}} \mathbb{E}^a_{\Pi_k(b)} \left(\left| \mathrm{val}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(B_1 \right) - \widetilde{\mathrm{val}}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(\Pi_k(B_1) \right) \right| \right) \\
&\leq \frac{\varepsilon}{T+1} + \max_{a \in \mathcal{A}} \mathbb{E}^a_{\Pi_k(b)} \left(\left| \mathrm{val}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(B_1 \right) - \widetilde{\mathrm{val}}^{t-1}_{\mathrm{BR}(\mathcal{X})} \left(\Pi_k(B_1) \right) \right| \right) && \text{(Definition of } k\text{)} \\
&\leq \frac{\varepsilon}{T+1} + \max_{a \in \mathcal{A}} \mathbb{E}^a_b (E_{t-1}(B_1)) && \text{(Definition of } E_{t-1}\text{)} \\
&\leq \frac{\varepsilon}{T+1} + \frac{t\varepsilon}{T+1} && \text{(Induction hypothesis)} \\
&= \frac{(t+1)\varepsilon}{T+1},
\end{aligned}$$

and the induction argument holds. Therefore, we obtain that $E_T(b) \leq \frac{(T+1)\varepsilon}{T+1} = \varepsilon$.

Complexity: For every action, the bellman update considers $O(|\mathcal{S}|^2|\mathcal{Z}|)$. Taking the maximum over actions, we get a complexity bound of $O(|\mathcal{S}|^2|\mathcal{A}||\mathcal{Z}|)$. We considered the update on a uniform grid $G_k \subseteq \Delta(\mathcal{S})$ with, for all $b \in \Delta(\mathcal{S})$, $\|b - \Pi_k(b)\|_1 \leq \varepsilon/(T+1)$ by taking $k = \left\lceil \frac{(T+1)|\mathcal{S}|}{\varepsilon} \right\rceil$. The size of grid in the $(|\mathcal{S}|-1)$-simplex $\Delta(\mathcal{S})$ is thus

$$|G_k| = \binom{k + |\mathcal{S}| - 1}{|\mathcal{S}| - 1} = O\left(\left(\tfrac{T|\mathcal{S}|}{\varepsilon} \right)^{|\mathcal{S}|-1} \right).$$

Therefore, we deduce that

$$\begin{aligned}\text{Total time complexity} &= O\left((T+1)\cdot|G_k|\cdot|\mathcal{S}|^2||\mathcal{A}||\mathcal{Z}|\right)\\ &= O\left((T+1)\cdot\left(\frac{T|\mathcal{S}|}{\varepsilon}\right)^{|\mathcal{S}|-1}\cdot|\mathcal{S}|^2||\mathcal{A}||\mathcal{Z}|\right)\\ &= O\left(\frac{T^{|\mathcal{S}|}|\mathcal{S}|^{|\mathcal{S}|+1}|\mathcal{A}||\mathcal{Z}|}{\varepsilon^{|\mathcal{S}|-1}}\right),\end{aligned}$$

which completes the proof. □

Theorem 1. *Quantitative analysis for belief-reachability objectives for revealing POMDPs is in* EXPTIME.

Proof. Let P be a revealing POMDP with initial belief $b \in \Delta(\mathcal{S})$ and set of target states $\mathcal{X} \subseteq \mathcal{S}$. By Lemma 8, there exists a (n,q)-stopping optimal policy σ with $n = |\mathcal{S}| + 2$ and $q = \delta_{\min}^2\left(\frac{\delta_{\min}}{|\mathcal{S}|}\right)^{|\mathcal{S}|}$. By Lemma 10, for every $\varepsilon > 0$, there exists a time horizon $T = \left\lceil \frac{n\log(\varepsilon)}{\log(1-q)} \right\rceil$ such that

$$\mathbb{P}_b^\sigma(\exists t \le T \quad B_t \in \mathcal{D}_\mathcal{X}) \ge \mathrm{val}_{\mathrm{BR}(\mathcal{X})}(b) - \varepsilon.$$

Moreover, by Lemma 11, we can compute an approximation of $\mathrm{val}^T_{\mathrm{BR}(\mathcal{X})}(b)$ denoted $\widetilde{\mathrm{val}}^T_{\mathrm{BR}(\mathcal{X})}$ with total time complexity

$$O\left(\frac{T^{|\mathcal{S}|}|\mathcal{S}|^{|\mathcal{S}|+1}|\mathcal{A}||\mathcal{Z}|}{\varepsilon^{|\mathcal{S}|-1}}\right).$$

We consider the grid G_k as previously defined in the proof of Lemma 11 with $k = \left\lceil \frac{(T+1)|\mathcal{S}|}{\varepsilon} \right\rceil$. For every $b \in \Delta(\mathcal{S})$, there exists $\Pi_k(b) \in G_k$ such that by taking $T = \left\lceil \frac{n\log(\varepsilon)}{\log(1-q)} \right\rceil$,

$$\begin{aligned}&\left|\mathrm{val}(b)_{\mathrm{BR}(\mathcal{X})} - \widetilde{\mathrm{val}}^T_{\mathrm{BR}(\mathcal{X})}\left(\Pi_k(b)\right)\right|\\ &\quad\le \left|\mathrm{val}(b)_{\mathrm{BR}(\mathcal{X})} - \mathbb{P}_b^\sigma(\exists t \le T \quad B_i \in \mathcal{D}_\mathcal{X})\right| + \left|\mathbb{P}_b^\sigma(\exists t \le T \quad B_t \in \mathcal{D}_\mathcal{X}) - \widetilde{\mathrm{val}}^T_{\mathrm{BR}(\mathcal{X})}\left(\Pi_k(b)\right)\right| \quad \text{(Triangle inequality)}\\ &\quad\le \varepsilon + \left|\mathbb{P}_b^\sigma(\exists t \le T \quad B_t \in \mathcal{D}_\mathcal{X}) - \widetilde{\mathrm{val}}^T_{\mathrm{BR}(\mathcal{X})}\left(\Pi_k(b)\right)\right| \quad \text{(Lemma 10)}\\ &\quad\le 2\varepsilon. \quad \text{(Lemma 11)}\end{aligned}$$

Let $L \in \mathbb{N}$ be the bit-length bound on transition probabilities so that $\delta_{\min} \ge 2^{-L}$. Since $q \in (0,1]$ and using that $\log(1-q) \le -q$, we have that

$$\begin{aligned}T &\le n\left(\frac{\log(\varepsilon)}{\log(1-q)} + 1\right)\\ &\le n\left(\frac{1}{q}\log\left(\frac{1}{\varepsilon}\right) + 1\right)\\ &\le n\left(\left(\frac{|\mathcal{A}|}{\delta_{\min}}\right)^{|\mathcal{S}|}\frac{1}{\delta_{\min}^2}\log\left(\frac{1}{\varepsilon}\right) + 1\right)\\ &\le n\left(\left(|\mathcal{A}|2^L\right)^{|\mathcal{S}|}2^{2L}\log\left(\frac{1}{\varepsilon}\right) + 1\right),\end{aligned}$$

and thus, we deduce the following upper bound for the horizon

$$T = O\left((|\mathcal{S}|+2)2^{L(|\mathcal{S}|+2)}|\mathcal{A}|^{|\mathcal{S}|}\log\left(\frac{1}{\varepsilon}\right)\right)$$

and thus, $T^{|\mathcal{S}|} = O\left((|\mathcal{S}|+2)^{|\mathcal{S}|}2^{L|\mathcal{S}|(|\mathcal{S}|+2)}|\mathcal{A}|^{|\mathcal{S}|^2}\log(1/\varepsilon)^{|\mathcal{S}|}\right)$. Inserting this bound into the above complexity yields

$$\text{Total time complexity} = O\left(\frac{(|\mathcal{S}|+2)^{|\mathcal{S}|}|\mathcal{S}|^{|\mathcal{S}|+1}|\mathcal{Z}|2^{L|\mathcal{S}|(|\mathcal{S}|+2)}|\mathcal{A}|^{|\mathcal{S}|^2+1}\log(1/\varepsilon)^{|\mathcal{S}|}}{\varepsilon^{|\mathcal{S}|-1}}\right).$$

Therefore, the total time complexity is upper bounded by $2^{\text{poly}(L,|\mathcal{S}|,\log(|\mathcal{A}|),\log(|\mathcal{Z}|),\log(1/\varepsilon))}$, which concludes the proof. □

B Proofs of Section 5

Lemma 13. *Consider a POMDP P and its underlying MDP M. For every reachable end component $\mathcal{U} = (\mathcal{Q},\mathcal{E})$ of M, i.e., there exists a policy σ on P such that $\mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = \mathcal{U}) > 0$, we have that, for all states $q \in \mathcal{Q}$, actions $a \in \mathcal{E}(q)$, and signals $z \in \mathcal{Z}$, there exists a policy $\sigma_{q,a,z}$ on P such that*

$$\mathbb{P}^{\sigma_{q,a,z}}_{b}(\mathcal{I}(\rho) \subseteq \mathcal{U}) \;=\; 1\,,$$

where b is the belief after starting with belief $\mathbb{1}[q]$, playing action a, and receiving signal z.

Proof. Consider a POMDP $P = (\mathcal{S},\mathcal{A},\mathcal{Z},\delta,b_0)$, its underlying MDP M with states $\mathcal{S}\times\mathcal{Z}$, an end component $\mathcal{U} = (\mathcal{Q},\mathcal{E})$ of M, and a policy σ on P such that $\mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = \mathcal{U}) > 0$. By contradiction, assume that there exists a state $q \in \mathcal{Q}$, an action $a \in \mathcal{E}(q)$, and a signal $z \in \mathcal{Z}$ such that, for all policies $\widetilde{\sigma}$ on P, we have

$$\mathbb{P}^{\widetilde{\sigma}}_{b}(\mathcal{I}(\rho) \subseteq \mathcal{U}) < 1\,, \tag{1}$$

where b is the belief after starting with belief $\mathbb{1}[q]$, playing action a, and receiving signal z. We show that $\mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = \mathcal{U}) = 0$, which is a contradiction.

By definition of end components, because $\mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = \mathcal{U}) > 0$, for every action $a \in \mathcal{E}(q)$, the policy σ plays action a infinitely often when the belief of the controller has positive probability on q. By Equation (1), for all policies $\widetilde{\sigma}$ on P, the set $\mathcal{S}\setminus\mathcal{Q}$ is reached with positive probability, i.e.,

$$\mathbb{P}^{\widetilde{\sigma}}_{b}(\exists t \geq 0 \quad S_t \in \mathcal{S}\setminus\mathcal{Q}) > 0\,. \tag{2}$$

We claim that the set $\mathcal{S}\setminus\mathcal{Q}$ is in fact reached within $2^{|\mathcal{Q}|}$ steps with positive probability. Formally, for all policies $\widetilde{\sigma}$,

$$\mathbb{P}^{\widetilde{\sigma}}_{b}(\exists t \leq 2^{|\mathcal{Q}|} \quad S_t \in \mathcal{S}\setminus\mathcal{Q}) > 0\,. \tag{3}$$

Indeed, we show that, starting from b, all policies reach a belief at which every action leads outside of $\mathcal{Q}$ with positive probability. Consider a belief $\widetilde{b} \subseteq \Delta(\mathcal{Q})$. Then, an action $a \in \mathcal{A}$ is safe for $\widetilde{b}$ if $\mathbb{P}^{a}_{\widetilde{b}}(\text{supp}(B_1) \subseteq \mathcal{Q}) = 1$. Note that, if an action is safe for a belief, then it is safe for all beliefs with the same support. Also, starting from b, if the controller can force the dynamic to visit only beliefs whose support has some safe action, then taking these actions one constructs a policy $\widetilde{\sigma}$ such that $\mathbb{P}^{\widetilde{\sigma}}_{b}(\forall t \geq 0 \quad S_t \in \mathcal{Q}) = 1$. By Equation (2), this is not the case, i.e., for all policies the dynamic must visit a belief whose support has no safe action. Moreover, for

each policy, such a belief is reached in the first $(2^{|\mathcal{Q}|}-1)$ steps because there are only so many supports, which proves our claim.

Equation (3) implies that the probability of reaching the set $\mathcal{S} \setminus \mathcal{Q}$ is lower bounded, i.e., for all policies $\widetilde{\sigma}$,

$$\mathbb{P}_b^{\widetilde{\sigma}}\left(\exists t \leq 2^{|\mathcal{Q}|} \quad S_t \in \mathcal{S} \setminus \mathcal{Q}\right) \;\geq\; (\delta_{\min})^{2^{|\mathcal{Q}|}} \,.$$

With this, we prove that $\mathbb{P}_{b_0}^{\sigma}(\mathcal{I}(\rho) = \mathcal{U}) = 0$, which is a contradiction.

Indeed, because $\mathbb{1}[q]$ is visited infinitely often, action a played infinitely often, and therefore the belief b is visited infinitely often. Conditioned on the event $\{\inf(\rho) = \mathcal{U}\}$, visits to the state q followed by playing the action a occur infinitely often under policy σ. The above inequality implies that some states $s \in \mathcal{S} \setminus \mathcal{Q}$ are reached with lower-bounded positive probability after such visits. Therefore, we get

$$\mathbb{P}_{b_0}^{\sigma}(\mathcal{I}(\rho) = \mathcal{U}) = 0\,,$$

which contradicts with the assumption, and therefore, yields the result. □

Lemma 14. *Consider a revealing POMDP P and its underlying MDP M. For every policy σ on P and end component $\mathcal{U} = (\mathcal{Q}, \mathcal{E})$ of M such that $\mathbb{P}_{b_0}^{\sigma}(\mathcal{I}(\rho) = \mathcal{U}) > 0$, we have, for all states $q \in \mathcal{Q}$, there exists a policy σ_q on P such that*

$$\mathbb{P}_{\mathbb{1}[q]}^{\sigma_q}(\mathcal{I}(\rho) = \mathcal{U}) \;=\; 1\,.$$

Proof. By Lemma 13, for all states $q \in \mathcal{Q}$, actions $a \in \mathcal{E}(q)$, and signals $z \in \mathcal{Z}$, there exists a policy $\sigma_{q,a,z}$ such that

$$\mathbb{P}_b^{\sigma_{q,a,z}}(\mathcal{I}(\rho) \subseteq \mathcal{U}) \;=\; 1\,,$$

where b is the belief after starting with belief $\mathbb{1}[q]$, playing action a, and receiving signal z. We now construct a policy σ_q such that

$$\mathbb{P}_{\mathbb{1}[q]}^{\sigma_q}(\mathcal{I}(\rho) = \mathcal{U}) \;=\; 1\,.$$

The policy proceeds in *phases*, one for every ordered pair of state $\widetilde{q} \in \mathcal{Q}$ and $\widetilde{a} \in \mathcal{E}(\widetilde{q})$. The goal of phase $(\widetilde{q}, \widetilde{a})$ is to reach the Dirac belief $\mathbb{1}[\widetilde{q}]$ and play action $\widetilde{a}$. By cycling through all pairs, each possible state–action pair is visited infinitely often. For each phase $(\widetilde{q}, \widetilde{a})$, the policy σ_q is as follows. While the current belief is *not* Dirac, σ_q follows the almost-sure safety policy defined in Lemma 13. Since the POMDP is revealing, a Dirac belief $\mathbb{1}[q']$ is reached almost-surely. Since $\mathcal{U}$ is an end component in the MDP, there exists a finite path $(q_0, a_0, q_1, a_1 \ldots, q_k)$ inside $\mathcal{U}$ such that $q_0 = q'$ and $q_k = \widetilde{q}$, and $q_{i+1} \in \mathrm{supp}(\delta(q_i, a_i))$. The policy now plays $(a_0, a_1, \ldots, a_{k-1})$ in order, stopping early if the belief becomes non-Dirac. In this case, it returns to the safety policy to stay inside $\mathcal{U}$. Once the belief is $\mathbb{1}[\widetilde{q}]$, σ_q plays the action $\widetilde{a}$ and then proceeds to the next phase. Therefore, all states and actions inside end component $\mathcal{U}$ are visited infinitely often, which completes the proof. □

Lemma 15. *Consider a revealing POMDP with parity objectives. Denote the set of states for which, if the initial belief were a Dirac on that state, then the parity condition can be satisfied almost-surely by $\mathcal{X} \coloneqq \left\{s \in \mathcal{S} : \exists \sigma \in \Sigma \quad \mathbb{P}_{\mathbb{1}[s]}^{\sigma}(\mathrm{Parity}) = 1\right\}$. Then, the parity value coincides with the belief-reachability value to $\mathcal{X}$, i.e., for all beliefs $b \in \Delta(\mathcal{S})$,*

$$\mathrm{val}_{\mathrm{P}}(b) = \mathrm{val}_{\mathrm{BR}(\mathcal{X})}(b)\,.$$

Proof. Consider a revealing POMDP P with the initial belief $b_0 \in \Delta(\mathcal{S})$. Denote the set of states for which, if the initial belief were a Dirac on that state, then the parity condition can be satisfied almost-surely by $\mathcal{X} \coloneqq \left\{ s \in \mathcal{S} : \exists \sigma \quad \mathbb{P}^{\sigma}_{\mathbb{1}[s]}(\text{Parity}) = 1 \right\}$. We show that, for all policies $\sigma \in \Sigma$,

$$\mathbb{P}^{\sigma}_{b_0}\big(\rho \in \text{Parity}\big) = \mathbb{P}^{\sigma}_{b_0}\big(\exists t \geq 0 \quad B_t \in \mathcal{D}_{\mathcal{X}}\big)\,,$$

which implies the equality by maximizing over the policy.

Fix a policy σ. Denote $\mathcal{W}$ the set of all end components of the underlying MDP on $\mathcal{S} \times \mathcal{Z}$ that can be reached by σ. Formally,

$$\mathcal{W} \coloneqq \{\mathcal{U} : \mathbb{P}^{\sigma}_{b_0}(\mathcal{I}(\rho) = U) > 0\}\,.$$

Fix an end component $\mathcal{U} = (\mathcal{Q}, \mathcal{E}) \in \mathcal{W}$. Note that, if a play ρ eventually remains in U forever, i.e., $\mathcal{I}(\rho) = \mathcal{U}$, then, by definition of end component, all the states $q \in \mathcal{Q}$ are visited infinitely often. Therefore, if $\mathcal{I}(\rho) = \mathcal{U}$, then

$$\rho \in \text{Parity} \iff \min\{\text{pri}(q) : q \in \mathcal{Q}\} \text{ is even.}$$

Consequently, the parity condition being satisfied depends on the end component reached. Formally, for all end components $\mathcal{U} \in \mathcal{W}$,

$$\mathbb{P}^{\sigma}_{b_0}\big(\rho \in \text{Parity} \mid \mathcal{I}(\rho) = \mathcal{U}\big) \in \{0, 1\}\,.$$

Denote by $\mathcal{W}_{\text{P}} \subseteq \mathcal{W}$ the set of end components where the parity condition is satisfied with probability 1 conditioned on eventually remaining in the end component. Formally,

$$\mathcal{W}_{\text{P}} \coloneqq \left\{\mathcal{U} \in \mathcal{W} : \mathbb{P}^{\sigma}_{b_0}\big(\rho \in \text{Parity} \mid \mathcal{I}(\rho) = \mathcal{U}\big) = 1\right\}\,.$$

Recall that, by Lemma 12, some end component must be reached. Fix an arbitrary end component $\mathcal{U} = (\mathcal{Q}, \mathcal{E}) \in \mathcal{W}_{\text{P}}$. By Lemma 14, for all states $q \in \mathcal{Q}$, there exists an almost-sure winning policy for the parity objective when the initial belief is $\mathbb{1}[q]$. Therefore, q is an almost-sure winning parity state. In conclusion, we get that

$$\begin{aligned}
\mathbb{P}^{\sigma}_{b_0}\big(\rho \in \text{Parity}\big) &= \sum_{\mathcal{U} \in \mathcal{W}_{\text{P}}} \mathbb{P}^{\sigma}_{b_0}\big(\mathcal{I}(\rho) = \mathcal{U}\big) && \text{(Lemma 12)} \\
&= \mathbb{P}^{\sigma}_{b_0}\big(\exists t \geq 0 \quad B_t \in \mathcal{D}_{\mathcal{X}}\big)\,, && \text{(Lemma 14)}
\end{aligned}$$

which yields the result. □

Theorem 3. *Quantitative analysis for parity objectives for revealing POMDPs is in* EXPTIME.

Proof. By [7, Theorem 3], computing almost-sure winning parity states in the revealing POMDP is in EXPTIME. By Theorem 1, approximating the belief-reachability value to this set of states is in EXPTIME. By Lemma 15, this value coincides with the parity value and therefore we computed an approximation of the parity value in EXPTIME. □

Theorem 4. *Optimal policies exist for parity objectives for revealing POMDPs.*

Proof. Let $\mathcal{S}_{\text{P}}$ be the set of almost-sure winning parity states. By Lemma 15, the parity value coincides with the belief-reachability value to the set $\mathcal{S}_{\text{P}}$. By Corollary 9, there exists an optimal belief-reachability policy σ_0. For all state $s \in \mathcal{S}_{\text{P}}$, let σ_s be the almost-sure winning parity policy when the initial belief is $\mathbb{1}[s]$. We now construct the optimal policy σ as follows. The optimal policy follows the policy σ_0 until it reaches a Dirac belief $\mathbb{1}[s]$ where $s \in \mathcal{S}_{\text{P}}$. Then, it switches to following the policy σ_s. The probability of satisfying the parity condition under the policy σ is the belief-reachability value to the set $\mathcal{S}_{\text{P}}$, which coincides with the parity value. This completes the proof. □

Corollary 5. *For revealing POMDPs with parity objectives, limit-sure and almost-sure winning coincide, and limit-sure analysis is in* EXPTIME-complete.

Proof. Consider revealing POMDPs with parity objectives. By Theorem 4, optimal policies exist. Therefore, if a revealing POMDP with parity objectives is limit-sure winning, then it is almost-sure winning. The reverse direction follows by definition. Lastly, by [7, Theorem 3], the existence of an almost-sure policy is in EXPTIME-complete. Therefore, limit-sure analysis is also EXPTIME-complete because the two problems coincide. □

C Outline of Algorithms

This section outlines three algorithms for quantitative analysis: for the T-step belief-reachability, for the belief-reachability, and for the parity value, respectively.

Given a belief $b \in \Delta(\mathcal{S})$, an action $a \in \mathcal{A}$, and a signal $z \in \mathcal{Z}$, recall that the belief update is defined by

$$\tau(b, a, z)(s') \coloneqq \frac{\sum_{s \in \mathcal{S}} b(s)\delta(s, a)(s', z)}{\sum_{s, s' \in \mathcal{S}} b(s)\delta(s, a)(s', z)}\,.$$

Outline of Algorithm 1 Given a revealing POMDP P, a target set $\mathcal{X} \subseteq \mathcal{S}$, a time horizon $T \in \mathbb{N}$, and an additive error $\varepsilon > 0$, the algorithm returns a value that is guaranteed to be within ε of the T-step belief-reachability value. Intuitively, Algorithm 1 replaces the continuous set of beliefs with a finite uniform grid to run value iteration only on those grid points.

The algorithm has two stages and follows from the proof of Lemma 11:

(1) We build a finite uniform grid of beliefs $G_k \subseteq \Delta(\mathcal{S})$ such that every belief is at most $\varepsilon/(T+1)$ away from some grid point.

(2) We perform T backward Bellman updates on the grid: each grid belief evaluates every action and keeps the one with the best expected value over the projected successor beliefs.

Because the T-step belief-reachability value is Lipschitz, the projection errors accumulated over the steps is at most ε. Therefore, the algorithm returns an epsilon approximation of the T-step belief-reachability value.

Outline of Algorithm 2 Given a revealing POMDP P, a target set $\mathcal{X} \subseteq \mathcal{S}$, and an additive error $\varepsilon > 0$, the algorithm return a value that is guaranteed to be within ε of the belief-reachability value. Intuitively, Algorithm 2 approximates the belief-reachability value by computing the T-step belief reachability value for a horizon T chosen large enough that the two differ by at most ε.

The algorithm proceeds as follows:

(1) We first choose T large enough. By Lemma 7 and Lemma 8, there is an (n, q)-stopping optimal policy σ with parameters $n = |\mathcal{S}| + 2$ and $q = \delta_{\min}^2(\delta_{\min}/|\mathcal{A}|)^{|\mathcal{S}|}$ for the belief-reachability objectives. Next, by Lemma 10, we set $T = n\lceil \log(\varepsilon/2)/\log(1-q)\rceil$ so that the T-step belief-reachability objective under σ is an $\varepsilon/2$ approximation of the belief-reachability value.

(2) We call Lemma 11 by taking $T = n\lceil \log(\varepsilon/2)/\log(1-q)\rceil$ and $\varepsilon/2$.

By Theorem 1, the algorithm returns an epsilon approximation of the belief-reachability value.

Outline of Algorithm 3 Given a revealing POMDP P, a priority function pri : $\mathcal{S} \to \{0, \ldots, d\}$, and an additive error $\varepsilon > 0$, the algorithm returns a value that is guaranteed to be within ε of the parity value. Intuitively, Algorithm 1 computes the approximation of the belief-reachability objectives to the set of Dirac on the almost-sure winning parity states.

The algorithm proceeds as follows:

(1) We use [7, Theorem 3] to compute the set of almost-sure winning states, denoted $\mathcal{S}_P$.

(2) We call Algorithm 2 by taking the target set as $\mathcal{S}_P$.

By Lemma 15, the parity value of P and the belief-reachability value to the set of Dirac on the almost-sure winning parity states objective coincide. Therefore, because Algorithm 2 returns an epsilon approximation of the belief reachability value, the algorithm returns an epsilon approximation of the parity value.

Algorithm 1: Approximation of T-step belief-reachability value

Input: Revealing POMDP $P = (\mathcal{S}, \mathcal{A}, \mathcal{Z}, \delta, b_0)$, target set $\mathcal{X} \subseteq \mathcal{S}$, horizon $T \in \mathbb{N}$, and additive error $\varepsilon > 0$

Output: v such that $\left| v - \mathrm{val}^T_{\mathrm{BR}(\mathcal{X})}(b_0) \right| \leq \varepsilon$

$k \leftarrow \left\lceil \frac{(T+1)|\mathcal{S}|}{\varepsilon} \right\rceil$;

$G_k \leftarrow \left\{ b \in \Delta(S) : b(s) = \frac{n(s)}{k},\ n(s) \in \{0, \ldots, k\},\ \sum_s n(s) = k \right\}$;

foreach $b \in G_k$ **do**

  $\widetilde{\mathrm{val}}^0_{\mathrm{BR}(\mathcal{X})}(b) \leftarrow \begin{cases} 1 & \text{if } b \in \mathcal{D}_{\mathcal{X}}, \\ 0 & \text{otherwise.} \end{cases}$;

end

for $t \leftarrow 1$ **to** T **do**

  foreach $b \in G_k$ **do**

    if $b \in \mathcal{D}_{\mathcal{X}}$ **then** $\widetilde{\mathrm{val}}^t_{\mathrm{BR}(\mathcal{X})}(b) \leftarrow 1$;

    else foreach $a \in \mathcal{A}$ **do**

      $Q_a \leftarrow 0$;

      foreach $z \in \mathcal{Z}$ **do**

        foreach $s, s' \in \mathcal{S}$ **do**

          $p \leftarrow b(s)\, \delta(s, a)(s', z)$;

          if $p > 0$ **then**

            $b' \leftarrow \Pi_k(\tau(b, a, z))$;

            $Q_a \leftarrow Q_a + p \cdot \widetilde{\mathrm{val}}^{t-1}_{\mathrm{BR}(\mathcal{X})}(b')$;

          end

        end

      end

    end

    $\widetilde{\mathrm{val}}^t_{\mathrm{BR}(\mathcal{X})}(b) \leftarrow \max_{a \in A} Q_a$;

  end

end

return $\widetilde{\mathrm{val}}^T_{\mathrm{BR}(\mathcal{X})}(\Pi_k(b))$;

Algorithm 2: Approximation of belief-reachability value

Input: Revealing POMDP $P = (\mathcal{S}, \mathcal{A}, \mathcal{Z}, \delta, b_0)$, target set $\mathcal{X} \subseteq \mathcal{S}$, and additive error $\varepsilon > 0$

Output: v such that $\left| v - \mathrm{val}_{\mathrm{BR}(\mathcal{X})}(b_0) \right| \leq \varepsilon$

$\delta_{\min} \leftarrow \min\{\delta(s, a)(s', z) : \forall s, s' \in \mathcal{S}, a \in \mathcal{A}, z \in \mathcal{Z} \quad \delta(s, a)(s', z) > 0\}$;

$n \leftarrow |\mathcal{S}| + 2$;

$T \leftarrow n \left\lceil \frac{\log(\varepsilon/2)}{\log(1-q)} \right\rceil$;

$v \leftarrow$ Algorithm 1$\big(P, \mathcal{X}, b_0, T, \varepsilon/2\big)$;

return v;

Algorithm 3: Approximation of parity value

Input: Revealing POMDP $P = (\mathcal{S}, \mathcal{A}, \mathcal{Z}, \delta, b_0)$, priority function $\mathrm{pri} : \mathcal{S} \to \{0, \ldots, d\}$, and additive error $\varepsilon > 0$

Output: v such that $|v - \mathrm{val}_{\mathrm{P}}(b_0)| \leq \varepsilon$

$\mathcal{X} \leftarrow$ AlmostSureParityStates(P, pri);

$v \leftarrow$ Algorithm 2$\big(P, \mathcal{X}, b_0, \varepsilon\big)$;

return v;