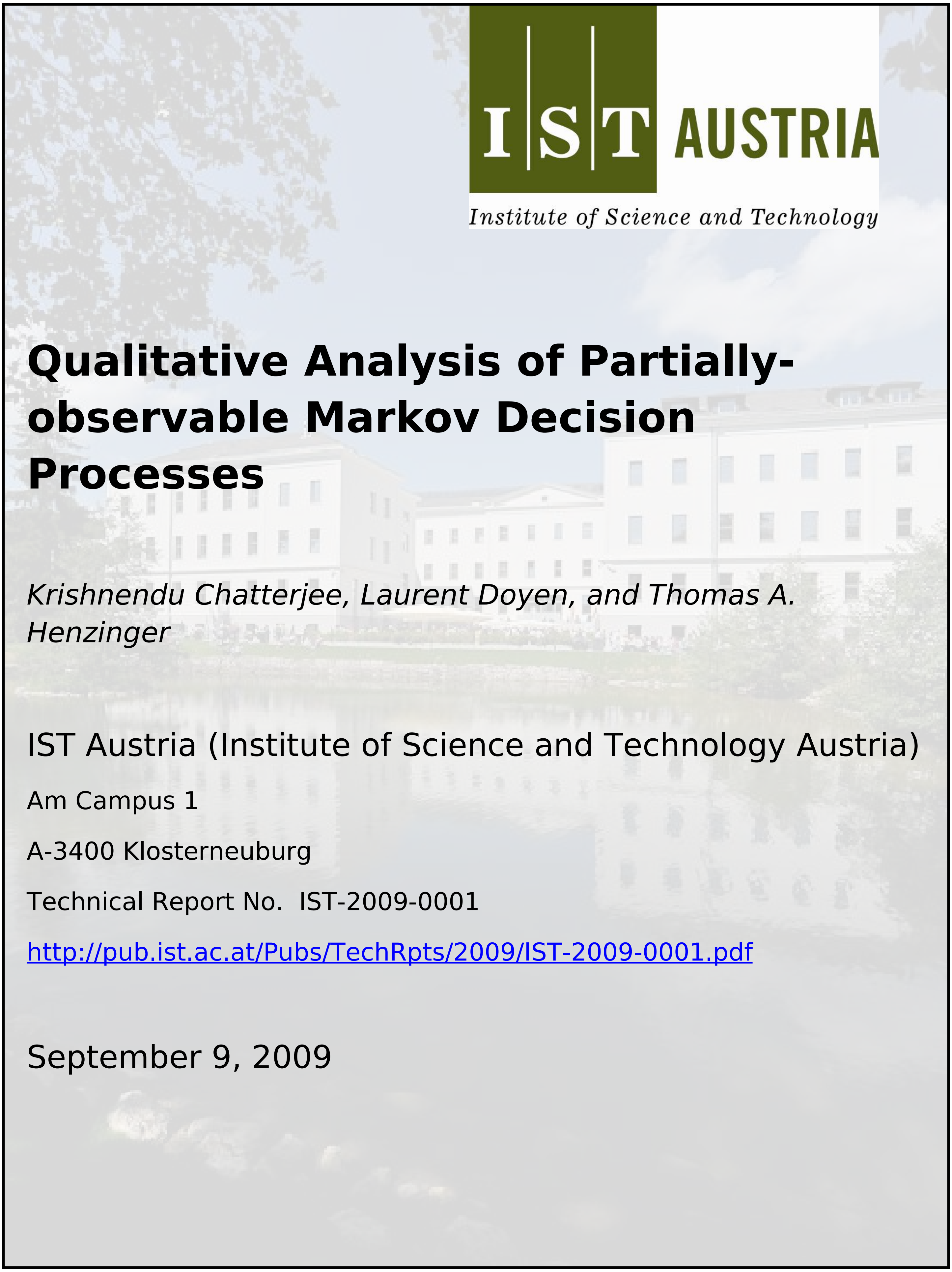

Qualitative Analysis of Partially-observable Markov Decision Processes

Krishnendu Chatterjee, Laurent Doyen, and Thomas A. Henzinger

IST Austria (Institute of Science and Technology Austria)

Am Campus 1

A-3400 Klosterneuburg

Technical Report No. IST-2009-0001

http://pub.ist.ac.at/Pubs/TechRpts/2009/IST-2009-0001.pdf

September 9, 2009

Qualitative Analysis of Partially-observable Markov Decision Processes

Krishnendu Chatterjee[1], Laurent Doyen[2], and Thomas A. Henzinger[1,3]

[1] IST Austria (Institute of Science and Technology Austria)
[2] CNRS, Cachan, France
[3] EPFL, Switzerland

Abstract. We study observation-based strategies for *partially-observable Markov decision processes (*POMDP*s)* with omega-regular objectives. An observation-based strategy relies on partial information about the history of a play, namely, on the past sequence of observations. We consider the qualitative analysis problem: given a POMDP with an omega-regular objective, whether there is an observation-based strategy to achieve the objective with probability 1 (almost-sure winning), or with positive probability (positive winning). Our main results are twofold. First, we present a complete picture of the computational complexity of the qualitative analysis of POMDPs with parity objectives (a canonical form to express omega-regular objectives) and its subclasses. Our contribution consists in establishing several upper and lower bounds that were not known in literature. Second, we present optimal bounds (matching upper and lower bounds) on the memory required by pure and randomized observation-based strategies for the qualitative analysis of POMDPs with parity objectives and its subclasses.

1 Introduction

Markov decision processes. *Markov decision processes (MDPs)* provide a model for systems that exhibit both probabilistic and nondeterministic behavior. MDPs were originally introduced to model and solve control problems for stochastic systems: there, nondeterminism represented the freedom in the choice of control action, while the probabilistic component of the behavior described the systems response to the control action. MDPs were later adopted as models for concurrent probabilistic systems, probabilistic systems operating in open environments [11], and under-specified probabilistic systems [3]

System specifications. The *specification* that describes the desired set of behaviors in the analysis of MDPs in verification of probabilistic systems and control of stochastic systems is typically an ω-regular set of paths in the MDP. A canonical way to express an ω-regular set of paths in MDPs is the classical parity objectives. The important sub-class of parity objectives include reachability, safety, Büchi (liveness) and coBüchi (co-liveness) objectives. Thus MDPs with parity objectives provides the theoretical framework to study important problems such as verification of probabilistic systems and control of stochastic systems.

Perfect vs. partial observations. MDPs can be broadly classified into two class: (a) MDPs with perfect observation; and (b) MDPs with partial observation. Most results about MDPs make the hypothesis of perfect observation. In this setting, the controller knows, during its interaction with the plant, the exact state of the plant. In practice, this hypothesis is often not reasonable. For example, in the context of hybrid systems, the controller acquires information about the state of the plant using sensors with finite precision, which return imperfect information about the state. Similarly, if the controller has only access to the public variables of the plant, not to their private variables, then the controller has only partial observation about the state of the plant. In these cases MDPs with partial observation is the more appropriate model.

Qualitative and quantitative analysis. The analysis of MDPs with parity objectives can be classified as the *qualitative analysis* and *quantitative analysis.* Given an MDP with an ω-regular specification the qualitative analysis asks for the computation of (a) the set of *almost-sure winning* states where the controller can satisfy the specification with probability 1; and (b) the set of *positive winning* states where the controller can satisfy the specification with positive probability. The more general quantitative analysis asks for the computation of the value at each state for the specification: the value at a state is the maximal probability with which the controller can satisfy the specification. MDPs with partial observation (POMDPs) are considerably more complicated than MDPs of perfect observation. First, decision problems for POMDPs usually lie in higher complexity classes than their perfect-observation counter-parts: for example, quantitative analysis for POMDPs with reachability objectives is undecidable whereas for MDPs with perfect observation the problem can be solved in polynomial time. Second, in the context of POMDPs witness winning strategies for qualitative analysis need memory even for simple objectives such as safety and reachability. This is again in contrast to the perfect-observation case, where memoryless strategies suffice for all parity objectives. Since the quantitative analysis of is undecidable for the simplest objectives such as reachability and safety for POMDPs, the qualitative analysis of POMDPs with parity objectives and its sub-classes is an important theoretical problem.

Our results. The contributions of this paper are twofold. First, we complete the picture for the complexity of qualitative analysis for POMDPs with parity objectives. Second, we present a complete characterization of the memory required by pure (deterministic) and randomized strategies for the qualitative analysis of POMDPs for the various sub-classes of parity objectives. We now present the details of our contribution. It was known from the results of [1] that almost-sure winning for reachability and Büchi objectives can be achieved in EXPTIME. It follows from the results of [1, 7] that the decision problem for almost-sure winning of coBüchi objectives and positive winning of Büchi objectives is undecidable. The EXPTIME-completeness for almost-sure winning for safety objectives follows from the results on games with partial observation. We show the following: (a) positive winning for reachability objectives is NLOGSPACE-complete; (b) positive winning for safety and coBüchi objectives can be achieved in EXPTIME; and (c) the almost-sure winning for reachability and positive winning for safety objectives is EXPTIME-hard (hence it follows that almost-sure winning for reachability and Büchi, and positive winning for safety and coBüchi objectives are EXPTIME-complete). This completes the picture for complexity of qualitative analysis for POMDPs with parity objectives. We present optimal memory bounds (matching upper and lower bound) for pure and randomized strategies for qualitative analysis: (a) for positive winning with reachability objectives randomized memoryless strategies suffice, and for pure strategies linear memory is necessary and sufficient; (b) for almost-sure winning of reachability and Büchi objectives, and for positive winning of safety and coBüchi objectives, and almost-sure winning for safety objectives we show that exponential memory is necessary and sufficient for both pure and randomized strategies. For positive winning of Büchi and almost-sure winning for coBüchi there is no bound on memory for strategies: this follows from the fact that these problems are undecidable.

Related work. Though MDPs has been more widely studied under the hypothesis of perfect observation, there are also several works that consider POMDPs. The work of [10, 9] considers POMDPs with several quantitative objectives for finite-horizon; whereas our work considers POMDPs with ω-regular objectives specified as parity objectives. The works of [1] presents several results related to the upper bound of qualitative analysis of POMDPs with parity objectives, and the works of [1, 7] shows undecidability results for some problems related to qualitative analysis of POMDPs with parity objectives. We present a solution to the remaining problems related to the qualitative analysis of POMDPs with parity objectives, and complete the picture.

2 Definitions

Notations. For a finite set A, a probability distribution on A is a function $\kappa : A \to [0, 1]$ such that $\sum_{a \in A} \kappa(a) = 1$. We denote the set of probability distributions on A by $\mathcal{D}(A)$. Given a distribution $\kappa \in \mathcal{D}(A)$, let $\mathsf{Supp}(\kappa) = \{a \in A \mid \kappa(a) > 0\}$ be the *support* of κ.

Games and Markov decision processes of partial observation. A *game structure* or a *Markov decision process (MDP)* (*of partial observation*) is a tuple $G = \langle L, \Sigma, \delta, \mathcal{O}, \gamma \rangle$, where L is a finite set of states, Σ is a finite set of actions, $\mathcal{O}$ is a finite set of observations, and $\gamma : \mathcal{O} \to 2^L \backslash \emptyset$ maps each observation to the set of states that it represents. In case of game structures, $\delta \subseteq L \times \Sigma \times L$ is a set of labeled transitions; and in case of MDPs $\delta : L \times A \to \mathcal{D}(L)$ is a probabilistic transition function. We require the following two properties on G: (i) for game structures we require that for all $\ell \in L$ and all $\sigma \in \Sigma$, there exists $\ell' \in L$ such that $(\ell, \sigma, \ell') \in \delta$; and (ii) for game structures and MDPs we require that the set $\{\gamma(o) \mid o \in \mathcal{O}\}$ partitions S. We refer to a game structure of partial observation as a POG and an MDP of partial observation as a POMDP. We say that G is a game structure or MDP of *perfect observation* if $\mathcal{O} = L$ and $\gamma(\ell) = \{\ell\}$ for all $\ell \in L$. We often omit $(\mathcal{O}, \gamma)$ in the description of games and MDPs of perfect observation. For a game structure G, for $\sigma \in \Sigma$ and $s \subseteq L$, let $\mathsf{Post}^G_\sigma(s) = \{\ell' \in L \mid \exists \ell \in s : (\ell, \sigma, \ell') \in \delta\}$; and for an MDP G, for $\sigma \in \Sigma$ and $s \subseteq L$, let $\mathsf{Post}^G_\sigma(s) = \{\ell' \in L \mid \exists \ell \in s : \delta(\ell, \sigma)(\ell') > 0\}$.

Plays. In a game structure, in each turn, Player 1 chooses an action in Σ, and Player 2 resolves nondeterminism by choosing the successor state, and in MDPs the successor state is chosen according to the probabilistic transition function. A *play* in G is an infinite sequence $\pi = \ell_0\sigma_0\ell_1 \dots \sigma_{n-1}\ell_n\sigma_n \dots$ such that for all $i \geq 0$, we have (a) $(\ell_i, \sigma_i, \ell_{i+1}) \in \delta$ if G is a game structure, and (b) $\delta(\ell_i, \sigma)(\ell_{i+1}) > 0$ if G is an MDP. The *prefix up to* ℓ_n of the play π is denoted by $\pi(n)$; its *length* is $|\pi(n)| = n + 1$; and its *last element* is $\mathsf{Last}(\pi(n)) = \ell_n$. The *observation sequence* of π is the unique infinite sequence $\gamma^{-1}(\pi) = o_0\sigma_0 o_1 \dots \sigma_{n-1} o_n \sigma_n \dots$ such that for all $i \geq 0$, we have $\ell_i \in \gamma(o_i)$. Similarly, the *observation sequence* of $\pi(n)$ is the prefix up to o_n of $\gamma^{-1}(\pi)$. The set of infinite plays in G is denoted $\mathsf{Plays}(G)$, and the set of corresponding finite prefixes is denoted $\mathsf{Prefs}(G)$. A state $\ell \in L$ is *reachable* in G if there exists a prefix $\rho \in \mathsf{Prefs}(G)$ such that $\mathsf{Last}(\rho) = \ell$. For a prefix $\rho \in \mathsf{Prefs}(G)$, the *cone* $\mathsf{Cone}(\rho) = \{\, \pi \in \mathsf{Plays}(G) \mid \rho \text{ is a prefix of } \pi \,\}$ is the set of plays that extend ρ. The *knowledge* associated with a finite observation sequence $\tau = o_0\sigma_0 o_1\sigma_1 \dots \sigma_{n-1} o_n$ is the set $\mathsf{K}(\tau)$ of states in which a play can be after this sequence of observations, that is, $\mathsf{K}(\tau) = \{\mathsf{Last}(\rho) \mid \rho \in \mathsf{Prefs}(G) \text{ and } \gamma^{-1}(\rho) = \tau\}$.

Lemma 1. *Let $G = \langle L, l_0, \Sigma, \delta, \mathcal{O}, \gamma \rangle$ be a* POG *or a* POMDP. *For $\sigma \in \Sigma$, $\ell \in L$, and $\rho, \rho' \in \mathsf{Prefs}(G)$ with $\rho' = \rho \cdot \sigma \cdot \ell$, let $o_\ell \in \mathcal{O}$ be the unique observation such that $\ell \in \gamma(o_\ell)$. Then $\mathsf{K}(\gamma^{-1}(\rho')) = \mathsf{Post}^G_\sigma(\mathsf{K}(\gamma^{-1}(\rho))) \cap \gamma(o_\ell)$.*

Strategies. A *pure strategy* in G for Player 1 is a function $\alpha : \mathsf{Prefs}(G) \to \Sigma$. A *randomized strategy* in G for Player 1 is a function $\alpha : \mathsf{Prefs}(G) \to \mathcal{D}(\Sigma)$. A (pure or randomized) strategy α for Player 1 is *observation-based* if for all prefixes $\rho, \rho' \in \mathsf{Prefs}(G)$, if $\gamma^{-1}(\rho) = \gamma^{-1}(\rho')$, then $\alpha(\rho) = \alpha(\rho')$. In the sequel, we are interested in the existence of observation-based strategies for Player 1. A *pure strategy* in G for Player 2 is a function $\beta : \mathsf{Prefs}(G) \times \Sigma \to L$ such that for all $\rho \in \mathsf{Prefs}(G)$ and all $\sigma \in \Sigma$, we have $(\mathsf{Last}(\rho), \sigma, \beta(\rho, \sigma)) \in \delta$. A *randomized strategy* in G for Player 2 is a function $\beta : \mathsf{Prefs}(G) \times \Sigma \to \mathcal{D}(L)$ such that for all $\rho \in \mathsf{Prefs}(G)$, all $\sigma \in \Sigma$, and all $\ell \in \mathsf{Supp}(\beta(\rho, \sigma))$, we have $(\mathsf{Last}(\rho), \sigma, \ell) \in \delta$. We denote by $\mathcal{A}_G$, $\mathcal{A}^O_G$, and $\mathcal{B}_G$ the set of all Player-1 strategies, the set of all observation-based Player-1 strategies, and the set of all Player-2 strategies in G, respectively.

Memory requirements of strategies. An equivalent definition of strategies is as follows. Let $\mathsf{M}em$ be a set called *memory.* An observation-based strategy with memory can be described as a pair of functions: (a) a *memory-update* function α_u: $\mathcal{O} \times \mathsf{M}em \to \mathsf{M}em$ that, given the memory and the current observation, updates the memory; and (b) a *next-action* function α_n: $\mathcal{O} \times \mathsf{M}em \to \mathcal{D}(\Sigma)$ that, given the memory and the observation, specifies the probability distribution of the next action (a pure strategy specifies the next action, rather than the probability distribution over actions). A strategy is *finite-memory* if the memory $\mathsf{M}em$ is finite and for a finite-memory strategy α the size of the strategy is the size of its memory, i.e., $|M|$. A strategy is *memoryless* if the memory $\mathsf{M}em$ is a singleton set. The memoryless strategies do not depend on the history of a play, but only on the current state. Each memoryless strategy for player 1 can be specified as a function $\alpha \colon \mathcal{O} \to \mathcal{D}(\Sigma)$.

Objectives. An *objective* for G is a set ϕ of infinite sequences of observations and actions, that is, $\phi \subseteq (\mathcal{O} \times \Sigma)^\omega$. A play $\pi = \ell_0\sigma_0\ell_1 \dots \sigma_{n-1}\ell_n\sigma_n \dots \in \mathsf{Plays}(G)$ *satisfies* the objective ϕ, denoted $\pi \models \phi$, if $\gamma^{-1}(\pi) \in \phi$. Objectives are generally Borel measurable: a Borel objective is a Borel set in the Cantor topology on $(\mathcal{O} \times \Sigma)^\omega$ [8]. We specifically consider reachability, safety, Büchi, coBüchi, and parity objectives, all of them Borel measurable. The parity objectives are a canonical form to express all ω-regular objectives [12]. For a play $\pi = \ell_0\sigma_0\ell_1 \dots$, we write $\mathsf{Inf}(\pi)$ for the set of observations that appear infinitely often in $\gamma^{-1}(\pi)$, that is, $\mathsf{Inf}(\pi) = \{o \in \mathcal{O} \mid \ell_i \in \gamma(o)$ for infinitely many i's$\}$.

- *Reachability and safety objectives.* Given a set $\mathcal{T} \subseteq \mathcal{O}$ of target observations, the *reachability* objective $\mathsf{Reach}(\mathcal{T})$ requires that an observation in $\mathcal{T}$ be visited at least once, that is, $\mathsf{Reach}(\mathcal{T}) = \{ \ell_0\sigma_0\ell_1\sigma_1 \dots \in \mathsf{Plays}(G) \mid \exists k \geq 0 \cdot \exists o \in \mathcal{T} : \ell_k \in \gamma(o)\}$. Dually, the *safety* objective $\mathsf{Safe}(\mathcal{T})$ requires that only observations in $\mathcal{T}$ be visited. Formally, $\mathsf{Safe}(\mathcal{T}) = \{ \ell_0\sigma_0\ell_1\sigma_1 \dots \in \mathsf{Plays}(G) \mid \forall k \geq 0 \cdot \exists o \in \mathcal{T} : \ell_k \in \gamma(o) \}$.
- *Büchi and coBüchi objectives.* The *Büchi* objective $\mathsf{Buchi}(\mathcal{T})$ requires that an observation in $\mathcal{T}$ be visited infinitely often, that is, $\mathsf{Buchi}(\mathcal{T}) = \{ \pi \mid \mathsf{Inf}(\pi) \cap \mathcal{T} \neq \emptyset \}$. Dually, the *coBüchi* objective $\mathsf{coBuchi}(\mathcal{T})$ requires that only observations in $\mathcal{T}$ be visited infinitely often. Formally, $\mathsf{coBuchi}(\mathcal{T}) = \{ \pi \mid \mathsf{Inf}(\pi) \subseteq \mathcal{T} \}$.
- *Parity objectives.* For $d \in \mathbb{N}$, let $p : \mathcal{O} \to \{0, 1, \dots, d\}$ be a *priority function*, which maps each observation to a nonnegative integer priority. The *parity* objective $\mathsf{Parity}(p)$ requires that the minimum priority that appears infinitely often be even. Formally, $\mathsf{Parity}(p) = \{ \pi \mid \min\{ p(o) \mid o \in \mathsf{Inf}(\pi) \}$ is even $\}$.

Observe that by definition, for all objectives ϕ, if $\pi \models \phi$ and $\gamma^{-1}(\pi) = \gamma^{-1}(\pi')$, then $\pi' \models \phi$. Given a Büchi objective $\mathsf{Buchi}(\mathcal{T})$ consider the priority function $p : \mathcal{O} \to \{ 0, 1 \}$ such that $p(o) = 0$ if $o \in \mathcal{T}$, and 1 otherwise; then we have $\mathsf{Parity}(p) = \mathsf{Buchi}(\mathcal{T})$. Similarly, given a coBüchi objective $\mathsf{coBuchi}(\mathcal{T})$ consider the priority function $p : \mathcal{O} \to \{1, 2\}$ such that $p(o) = 2$ if $o \in \mathcal{T}$, and 1 otherwise; then we have $\mathsf{Parity}(p) = \mathsf{coBuchi}(\mathcal{T})$. Hence Büchi and coBüchi objectives are special cases of parity objectives with two priorities.

Almost-sure and positive winning. An *event* is a measurable set of plays, and given strategies α and β for the two players (resp. a strategy α for Player 1 in MDPs), the probabilities of events are uniquely defined [13]. For a Borel objective ϕ, we denote by $\mathrm{Pr}_\ell^{\alpha,\beta}(\phi)$ (resp. $\mathrm{Pr}_\ell^{\alpha}(\phi)$ for MDPs) the probability that ϕ is satisfied from the starting state ℓ given the strategies α and β (resp. given the strategy α). Given a game structure G and a state ℓ, a strategy α for Player 1 is *almost-sure winning (almost winning in short)* (resp. *positive winning*) for the objective ϕ from ℓ if for all randomized strategies β for Player 2, we have $\Pr_{\ell}^{\alpha,\beta}(\phi) = 1$ (resp. $\Pr_{\ell}^{\alpha,\beta}(\phi) > 0$). Given an MDP G and a state ℓ, a strategy α for Player 1 is almost winning (resp. positive winning) for the objective ϕ from ℓ if we have $\Pr_{\ell}^{\alpha}(\phi) = 1$ (resp. $\Pr_{\ell}^{\alpha}(\phi) > 0$). We are interested in the decision problems of existence of observation-based strategies for Player 1 that is almost winning (resp. positive winning) from a given state ℓ.

3 Upper Bounds for the Qualitative Analysis of POMDPs

In this section we present upper bounds for the qualitative analysis of POMDPs. It follows from the results of [1] that the decision problems for almost winning for POMDPs with reachability, safety, and Büchi objectives can be solved in EXPTIME. It also follows from the results of [1] that the decision problem for almost winning for coBüchi objectives is undecidable if the strategies are restricted to be pure, and the results of [7] shows that the problem remains undecidable even if randomized strategies are considered. It also from the above results that the decision problem for positive winning in POMDPs with Büchi objectives is undecidable. In this section we present upper bounds for the decision problems for positive winning for safety, reachability and coBüchi objectives to complete the results on upper bounds on qualitative analysis of POMDPs.

3.1 Positive winning for reachability objectives

We first argue that the decision problem for positive winning with reachability objectives in POMDPs is NLOGSPACE-complete.

Reduction to graph reachability. Given a POMDP $G = \langle L, \Sigma, \delta, \mathcal{O}, \gamma \rangle$ consider the graph $\overline{G} = \langle L, E \rangle$ as follows: $(\ell, \ell') \in E$ if there exists an action $\sigma \in \Sigma$ such that $\delta(\ell, \sigma)(\ell') > 0$. Let $\mathcal{T} \subseteq \mathcal{O}$ be the set of target observations, and let $T = \bigcup_{o \in \mathcal{T}} \gamma(o)$ be the set of states that belong to the target observation. Let ℓ be a starting state, then the following assertions hold: (a) if there is a path π in $\overline{G}$ that reaches a state $t \in T$, then the randomized memoryless strategy for Player 1 in G that plays all actions uniformly at random ensures that the path π is executed in G with positive probability (i.e., ensure positive winning for $\mathsf{Reach}(\mathcal{T})$ in G from ℓ); and (b) if there is no path in $\overline{G}$ to reach T from ℓ, then there is no strategy (and hence no observation-based strategy) for Player 1 in G to achieve $\mathsf{Reach}(\mathcal{T})$. This shows that positive winning in POMDPs can be decided in NLOGSPACE. Graphs are a special case of POMDPs and hence graph reachability can be reduced to reachability with positive probability in POMDPs. Hence it follows that the decision problem for positive winning with reachability objectives is NLOGSPACE-complete.

Theorem 1. *Given a POMDP G with a reachability objective and a starting state ℓ, the decision problem that whether there is a positive winning strategy from ℓ is NLOGSPACE-complete.*

3.2 Positive winning for safety and coBüchi objectives

In this section we show that the decision problem for positive winning with safety and coBüchi objectives for POMDPs can be solved in EXPTIME. We first present the result for safety objectives, and the result is based on the subset (or knowledge-based) construction.

Subset construction. Given a POMDP $G = \langle L, \Sigma, \delta, \mathcal{O}, \gamma \rangle$, we define the *knowledge-based subset construction* of G as the MDP of perfect observation:

$$G^{\mathsf{K}} = \langle \mathcal{L}, \Sigma, \delta^{\mathsf{K}} \rangle,$$

where $\mathcal{L} = 2^L \backslash \{\emptyset\}$, and $\delta^{\mathsf{K}}(s_1, \sigma)(s_2) > 0$ iff there exists an observation $o \in \mathcal{O}$ such that $s_2 = \mathsf{Post}^G_\sigma(s_1) \cap \gamma(o)$ and $s_2 \neq \emptyset$. Moreover, all transition positive probabilities are assigned such that the probability distribution is uniform over its support. We refer to states in G^{K} as cells. A (pure or randomized) strategy in G^{K} is called a *knowledge-based* strategy. To distinguish between a general strategy in G, an observation-based strategy in G, and a knowledge-based strategy in G^{K}, we often use the notations α, α^o, and α^{K}, respectively.

Lemma 2. *For all cells $s \in \mathcal{L}$ that are reachable in G^{K} from a starting location $s' \subseteq \gamma(o')$ for some observation $o' \in \mathcal{O}$, for all observations $o \in \mathcal{O}$, either $s \subseteq \gamma(o)$ or $s \cap \gamma(o) = \emptyset$.*

By an abuse of notation, we define the *observation sequence* of a play $\pi = s_0\sigma_0 s_1 \dots \sigma_{n-1} s_n \sigma_n \dots \in \mathsf{Plays}(G^{\mathsf{K}})$ as the infinite sequence $\gamma^{-1}(\pi) = o_0\sigma_0 o_1 \dots \sigma_{n-1} o_n \sigma_n \dots$ of observations such that for all $i \geq 0$, we have $s_i \subseteq \gamma(o_i)$. This sequence is unique by Lemma 2. The play π *satisfies* an objective $\phi \subseteq (\mathcal{O} \times \Sigma)^\omega$ if $\gamma^{-1}(\pi) \in \phi$. Given a POMDP G with a target set $\mathcal{T}$ of observations, and the safety objective $\mathsf{Safe}(\mathcal{T})$ without loss of generality we assume that every state $\ell \in L$ such that $\gamma^{-1}(\ell) \notin \mathcal{T}$ is an absorbing state (i.e., state with only self-loops as out-going transitions): we assume so because if a play reaches a state with an observation not in $\mathcal{T}$, then the play is anyway loosing for Player 1.

Lemma 3. *Consider a* POMDP *G and the MDP G^{K} constructed by subset construction. Let $\mathcal{T}$ be the set of target observations, and let $F = \{ s \subseteq L \mid s \subseteq \gamma(o), o \in \mathcal{T} \}$. Let W be the set of cells in G^{K} such that Player 1 has a positive winning strategy for the objective* $\mathsf{Safe}(F)$*. The following assertions hold:*

1. *If the initial knowledge for Player* 1 *is a cell in W in G, then there is an observation-based strategy in G to satisfy* $\mathsf{Safe}(\mathcal{T})$ *with positive probability.*
2. *If the initial knowledge for Player 1 is a cell in $2^L \setminus W$, then there is no observation-based strategy for Player 1 to satisfy* $\mathsf{Safe}(\mathcal{T})$ *with positive probability.*

Proof. We present the proof in two parts. WLOG we assume that every state with observations not in $\mathcal{T}$ is absorbing, and hence cells in $2^L \setminus F$ are also absorbing.

1. Since G^{K} is an MDP of perfect observation, it follows that if there is a positive winning strategy for the safety objective $\mathsf{Safe}(F)$, then there is a pure memoryless strategy in G^{K} that is also positive winning for $\mathsf{Safe}(F)$ [6, 5]. Let α^{K} be a pure memoryless positive winning strategy for Player 1 in G^{K} for the objective $\mathsf{Safe}(F)$ from the cells in W. Define α^o a strategy for Player 1 in G as follows: for every $\rho \in \mathsf{Prefs}(G)$, let $\alpha^o(\rho) = \alpha^{\mathsf{K}}(\rho^{\mathsf{K}})$ where ρ^{K} is defined from $\rho = \ell_0\sigma_0\ell_1 \dots \sigma_{n-1}\ell_n$ by $\rho^{\mathsf{K}} = s_0\sigma_0 s_1 \dots \sigma_{n-1}s_n$ where $s_i = \mathsf{K}(\gamma^{-1}(\ell_0\sigma_0\ell_1 \dots \sigma_{i-1}\ell_i))$ for each $0 \leq i \leq n$. Clearly, α^o is a pure observation-based strategy as $\gamma^{-1}(\rho) = \gamma^{-1}(\rho')$ implies $\rho^{\mathsf{K}} = \rho'^{\mathsf{K}}$. Since the strategy α^{K} is pure memoryless in G^{K}, the strategy α^o is a finite-memory pure strategy with at most exponential ($2^{O(|L|)}$) memory. Once the strategy α^{K} is fixed in G^{K} we obtain a Markov chain. Since α^{K} is positive winning, it follows that in the Markov chain obtained, from every cell in W a closed recurrent set $C \subseteq W \subseteq F$ is reached with positive probability. Hence if the strategy α^o constructed from α^{K} is fixed in G and the initial knowledge is a cell in W, then it ensures the following: (a) there exist a subset O of states in L that are labeled by observations in $\mathcal{T}$, and once the set O is reached, then the set O is never left (this corresponds to the closed recurrent set C in G^{K}); and (b) with positive probability a path is executed that goes through only states labeled by observations in $\mathcal{T}$, and reaches the set O. Hence the strategy α^o is positive winning in G, given the initial knowledge is a cell in W.
2. Let $\overline{W} = 2^L \setminus W$ be the set of cells in G^{K} such that there is no positive winning strategy for the objective $\mathsf{Safe}(F)$. Hence from any cell in $\overline{W}$, for any knowledge-based strategy for Player 1 the following property hold: the play remains in $\overline{W}$ and reaches cells in $\overline{W} \setminus F$ with some positive probability $\eta > 0$ in $2^{|L|}$ steps. It follows if the initial knowledge for Player 1 in G is a cell in $\overline{W}$, then for any observation-based strategy in G, the probability to reach an observation in $\mathcal{O} \setminus \mathcal{T}$ in $k \cdot 2^{|L|}$ steps is at least $1-(1-\eta')^k$, for some $\eta' > 0$. As k goes to ∞, the value of $1-(1-\eta')^k$ goes to 1. Hence the probability to stay safe in $\mathcal{T}$ is 0 for any observation-based Player 1 strategy, with the initial knowledge in $\overline{W}$.

The result follows. ■

In the case of POMDPs, given the starting state is a state $\ell \in L$, the initial knowledge is the cell $\gamma^{-1}(\ell)$. Hence the decision problem for positive winning in POMDPs with safety objectives can be solved by solving the same problem for MDPs with perfect observation of exponential size. Since positive winning in MDPs of perfect observation with safety objectives can be solved in polynomial time [6, 5], we obtain an EXPTIME upper bound for POMDPs.

Theorem 2. *Given a* POMDP *G with a safety objective and a starting state ℓ, the decision problem that whether there is a positive winning strategy from ℓ can be decided in EXPTIME.*

Positive winning for coBüchi objectives. We now show that we can solve positive winning for POMDPs with coBüchi objectives by iteratively solving for positive winning with reachability and safety objectives in POMDPs. Let G be a POMDP with a coBüchi objective $\mathsf{coBuchi}(\mathcal{T})$, where $\mathcal{T} \subseteq \mathcal{O}$. We construct the MDP G^{K} of perfect observation, and let $C = \{ s \subseteq L \mid s \subseteq \gamma(o), o \in \mathcal{T} \}$. We consider the positive winning for the coBüchi objective $\mathsf{coBuchi}(C)$ in G^{K}. The set W of positive winning cells in G^{K} is obtained as follows:

1. let $W_0 = \emptyset$;
2. we obtain W_{i+1} from W_i as follows: let Z_i be the set of cells in G^{K} such that Player 1 can ensure staying safe in $C \cup W_i$ with positive probability, and W_{i+1} is obtained as the set of states that can reach Z_i with positive probability.

It follows from above that if the current knowledge is a cell in W_{i+1}, then either W_i is reached with positive probability or the play eventually only visits states in C. It follows from the proof of correctness for positive winning in POMDPs with safety objective, that if the current knowledge is a cell in W_{i+1}, then there is an observation-based strategy for Player 1 to ensure $\mathsf{coBuchi}(\mathcal{T})$. Let W be the fixpoint of the iteration, i.e., for some k we have $W_k = W_{k+1} = W$. Let $\overline{W} = 2^L \setminus W$. Then the following assertions hold.

1. From any cell $\overline{W}$, Player 1 cannot ensure positive probability to stay safe in $C \cup W$. Otherwise, such a cell would have been included in Z_{k+1} and hence it would contradict that $W_k = W_{k+1}$. Hence for every Player 1 knowledge-based strategy, if the initial knowledge is a cell in $\overline{W}$, then the set $(2^L \setminus C) \cap \overline{W}$ is reached with probability 1. It follows that if the current knowledge is a cell in $\overline{W}$, then for any observation-based strategy in G, observations in the set $\mathcal{O} \setminus \mathcal{T}$ is reached with probability 1.
2. From every cell $\overline{W}$, Player 1 cannot ensure to reach W with positive probability. Hence for every knowledge-based strategy for Player 1, if the initial knowledge is a cell in $\overline{W}$, then the play stays safe in $\overline{W}$ with probability 1. Hence given a knowledge-based strategy for Player 1, from every cell in $\overline{W}$, the set $(2^L \setminus C) \cap \overline{W}$ is reached with probability 1 and the game stays safe in $\overline{W}$. It follows that the set $(2^L \setminus C) \cap \overline{W}$ is visited with infinitely often with probability 1, and this ensures that from $\overline{W}$ the coBüchi condition is falsified with probability 1. Hence if the initial knowledge is a cell in $\overline{W}$, then for any observation-based strategy in G, the coBüchi objective $\mathsf{coBuchi}(\mathcal{T})$ is falsified with probability 1.

Hence by iteratively solving positive winning in POMDPs with reachability and safety objectives, the positive winning in POMDPs with coBüchi objectives can be achieved. Hence we have the following result.

Theorem 3. *Given a* POMDP G *with a coBüchi objective and a starting state* ℓ*, the decision problem that whether there is a positive winning strategy from* ℓ *can be decided in EXPTIME.*

4 Lower Bounds for the Qualitative Analysis of POMDPs

In this section we present lower bounds for the qualitative analysis of POMDPs. We first present the lower bounds for MDPs with perfect observation.

4.1 Lower bounds for MDPs with perfect observations

In the previous section we argued that for reachability objectives even in POMDPs that positive winning problem can be solved in NLOGSPACE. The lower bound of NLOGSPACE follows from a simple reduction of the reachability problem in graphs. For safety objectives and almost winning it is known that an MDP can be equivalently considered as a game where Player 2 makes choices of the successors from the support of the probability distribution of the transition function, and the almost winning set is same in the MDP and the game. Similarly, there is a reduction of games of perfect observations to MDPs of perfect observation for almost winning with safety objectives. Since the problem of almost winning in games of perfect observation is PTIME-complete, the result follows. We now show that the almost winning problem for reachability and the positive winning problem for safety objectives is PTIME-complete for MDPs with perfect observation.

Reduction from Circuit-Value-Problem **(CVP).** The CVP is as follows: let $N = \{ 1, 2, \ldots, n \}$ be a set of AND and OR gates, and I be a set of inputs. The set of inputs is partitioned into I_0 and I_1; I_0 is the set of inputs set to 0 (false) and I_1 is the set of inputs set to 1 (true). Every gate receives two inputs and produces an output; the inputs of a gate are outputs of another gate or an input from the set I. The connection graph of the circuit must be acyclic. Let the gate represented by the node 1 be the output node. The problem of deciding whether the output is 1 or 0 is PTIME-complete. We now present reduction of the CVP to MDPs with perfect observation for almost winning with reachability, and positive winning with safety objectives.

1. *Almost reachability.* Given the CVP, we construct the MDP of perfect observation as follows: (a) the set states is $N \cup I$; (b) the action set is $\Sigma = \{ l, r \}$; (c) the transition function is as follows: every node in I is absorbing, and for a state that represents a gate, (i) if it is a OR gate, then for the action l the left input gate is chosen with probability 1, and for the action r the right input gate is chosen with probability 1; and (ii) if it is AND gate, then irrespective of the action the left and right input gate is chosen with probability 1/2. The output of the CVP from node 1 is 1 iff the set I_1 is reached from the state 1 in the MDP with probability 1 (i.e., the state 1 is almost winning for the reachability objective $\mathsf{Reach}(I_1)$.)
2. *Positive safety.* For positive winning with safety objectives, we take the CVP, apply the same reduction as the reduction for almost reachability with the following modifications: every state in I_0 remains absorbing and from every state in I_1 the next state is the starting state 1 with probability 1 irrespective of the action. The set of safety target is the set $I_1 \cup N$. If the output of the

CVP problem is 1, then from the starting state the set I_1 is reached with probability 1, and hence the safety objective with the target $N \cup I_1$ is ensured with probability 1. If the output of the CVP problem is 0, then from the starting state the set I_0 is reached with positive probability $\eta > 0$ in n steps against all strategies. Since from every state in I_1 the successor state is the state 1, it follows that the probability to reach I_0 from the starting state 1 in $k \cdot (n+1)$ steps is at least $1 - (1-\eta)^k$, and this goes to 1 as k goes to ∞. Hence it follows that from state 1, the answer to the positive winning for the safety objective $\mathsf{Safe}(N \cup I_1)$ is YES iff the output to the CVP is 1.

Theorem 4. *Let G be an MDP of perfect observation, and let T be a subset of states. Whether the set T can be reached with positive probability is NLOGSPACE-complete; and whether the set T can reached with probability 1 or whether the safety in the set T can be ensured with probability 1 or positive probability is PTIME-complete.*

4.2 Lower bounds for POMDPs

We have already shown that positive winning with reachability objectives in POMDPs is NLOGSPACE-complete. As in the case of MDPs with perfect observation, for safety objectives and almost winning a POMDP can be equivalently considered as a game of partial observation where Player 2 makes choices of the successors from the support of the probability distribution of the transition function, and the almost winning set is same in the POMDP and the game. Since the problem of almost winning in games of partial observation with safety objective is EXPTIME-complete [2], the EXPTIME-completeness result follows. We now show that almost winning with reachability objectives and positive winning with safety objectives is EXPTIME-complete. Before the result we first present a discussion on Alternating Polynomial-space (PSPACE) Turing Machines (APTM).
Discussion. Let M be APTM and w be a input word, then there is an exponential bound on the number of configuration states. Hence if M can accept the word w, then it can be done within some $k_{|w|}$ steps, here $|w|$ is the length of the word w, and $k_{|w|}$ is bounded by exponential in $|w|$. We construct an equivalent APTM M' that behaves as M but keeps track (in polynomial space) in a counter the number of steps of M, and given a word $|w|$, if the number of steps crosses $k_{|w|}$ without accepting, then the word is rejected. The machine M' is equivalent to M and reaches the accepting or rejecting states in a number of steps bounded by an exponential in the length of the input word. The problem of given an APTM M and a word w, whether M accepts w is EXPTIME-complete.

Lower bounds. Let M be an APTM such that for every input word w, the accepting or the rejecting state is reached within exponential steps in $|w|$. A polynomial-time reduction R_G of an APTM M and an input word w to a game structure G of partial observation is given in [4] such that (a) $R_G(M, w) = G$; (b) there is a special accepting state in G; (c) M accepts w iff there is a observation-based strategy for Player 1 in G to reach the accepting state with probability 1. If the above reduction is applied to M, then the game structure satisfies the following additional properties: there is a special rejecting state that is absorbing, and for every observation-based strategy for Player 1, either (a) against all Player 2 strategies the accepting state is reached with probability 1; or (b) there is a pure Player 2 strategy that reaches the rejecting state with positive probability $\eta > 0$ in $2^{|L|}$ steps and the accepting or the rejecting state is reached with probability 1 in $2^{|L|}$ steps. We now present the reduction to POMDPs:

1. *Almost winning with reachability.* Given APTM M and w, let $G = R_G(M, w)$. We construct a POMDP G' from G as follows: we only modify the transition function in G' by uniformly choosing over the successor choices. Formally, for a state $\ell \in L$ and an action $\sigma \in \Sigma$ the probabilistic transition function δ' in G' is as follows:

$$\delta'(\ell, \sigma)(\ell') = \begin{cases} 0 & (\ell, \sigma, \ell') \notin \delta; \\ 1/|\{\, \ell_1 \mid (\ell, \sigma, \ell_1) \in \delta \,\}| & (\ell, \sigma, \ell') \in \delta. \end{cases}$$

   Given an observation-based strategy for Player 1 in G, we consider the same strategy in G': (1) if the strategy the reaches accepting state with probability 1 against all Player 2 strategies in G, then the strategy ensures that in G' the accepting state is reached with probability 1; and (2) otherwise there is a pure Player 2 strategy β in G that ensures the rejecting state is reached in $2^{|L|}$ steps with probability $\eta > 0$, and with probability at least $(1/|L|)^{|L|}$ the choices of the successors of strategy β is chosen in G', and hence the rejecting state is reached with probability at least $(1/|L|)^{|L|} \cdot \eta > 0$. It follows that in G' there is an observation-based strategy for almost winning the reachability objective with target of the accepting state iff there is such a strategy in G. The result follows.
2. *Positive winning with safety.* The reduction is same as above. We obtain the POMDP G'' from the POMDP G' above by making the following modification: from the state accepting, the POMDP goes back to the initial state with probability 1. If there is an observation-based strategy α for Player 1 in G' to reach the accepting state, then repeating the strategy α everytime the accepting state is visited, it can be ensured that the rejecting state is reached with probability 0. Otherwise, against every observation-based strategy for Player 1, the probability to reach the rejecting state in $k \cdot (2^{|L|} + 1)$ steps is at least $1 - (1 - \eta')^k$, where $\eta' = \eta \cdot (1/|L|)^{|L|} > 0$ (this is because there is a probability to reach the rejecting state with probability at least η' in $2^{|L|}$ steps, and unless the rejecting state is reached the starting state is again reached within $2^{|L|} + 1$ steps). Hence the probability to reach the rejecting state is 1. It follows that G' is almost winning for the reachability objective with the target of the accepting state iff in G'' there is an observation-based strategy for Player 1 to ensure that the rejecting state is avoided with positive probability. This completes the proof of correctness of the reduction.

Hence we have the following theorem, and the results are summarized in Table 1.

Theorem 5. *Let G be a* POMDP *and $\mathcal{T}$ be a set of observations. Whether the set $\mathcal{T}$ can be reached with positive probability is NLOGSPACE-complete; and whether the set $\mathcal{T}$ can reached with probability 1 or whether the safety in the set $\mathcal{T}$ can be ensured with probability 1 or positive probability is EXPTIME-complete.*

	Positive	Almost
Reachability	NLOGSPACE-complete (up+lo)	EXPTIME-complete (lo)
Safety	EXPTIME-complete (up+lo)	EXPTIME-complete
Büchi	Undecidable	EXPTIME-complete (lo)
coBüchi	EXPTIME-complete (up+lo)	Undecidable
Parity	Undecidable	Undecidable

Table 1. Computational complexity of POMDPs with different classes of parity objectives for positive and almost winning. Our contribution of upper and lower bounds are indicated as "up" and "lo" respectively in parenthesis.

5 Optimal Memory Bounds for Strategies

In this section we present optimal bounds on the memory required by pure and randomized strategies for positive and almost winning for reachability, safety, Büchi and coBüchi objectives.

5.1 Bounds for safety objectives

In this subsection we present optimal memory bounds for strategies for positive and almost winning with safety objectives in POMDPs. It follows from the correctness argument of Theorem 2 (the proof of Lemma 3) the pure strategies with exponential memory is sufficient for positive winning with safety objectives in POMDPs, and the exponential upper bound on memory of pure strategies for almost winning with safety objectives in POMDPs follows from the reduction to games. We now present a matching exponential lower bound for randomized strategies.

Lemma 4. *There exists a family $(P_n)_{n\in\mathbb{N}}$ of* POMDP*s of size $O(p(n))$ for a polynomial p with a safety objective such that the following assertions hold: (a) Player 1 has an almost (and therefore also positive) winning strategy in each of these* POMDP*s; and (b) there exists a polynomial q such that every finite-memory randomized strategy for Player 1 that is positive (or almost) winning in P_n has at least $2^{q(n)}$ states.*

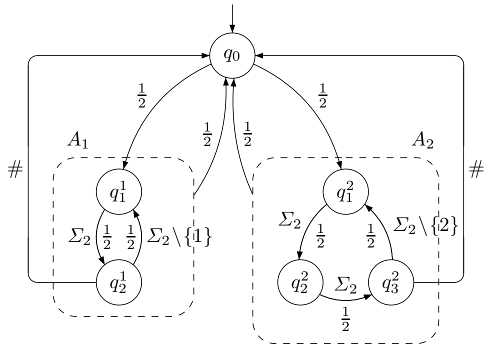

Fig. 1. The POMDP P_2.

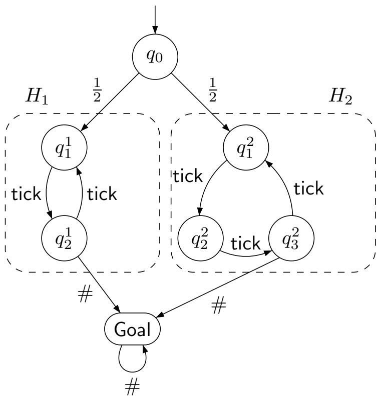

Fig. 2. The POMDP P_2'.

Preliminary. Let $p_1, p_2, \ldots$ be the list of prime numbers in increasing order. For $n \geq 1$, let $\Sigma_n = \{1, \ldots, n\}$. The set of actions of the POMDP P_n is $\Sigma_n \cup \{\#\}$. The POMDP is composed of an initial state q_0 and n sub-MDPs A_i, each consisting of a loop over p_i states $q_1, \ldots, q_{p_i}$. From each state q_j $(1 \leq j < p_i)$, every action in Σ_n leads to the next state q_{j+1} with probability $\frac{1}{2}$, and to the initial state q_0 with probability $\frac{1}{2}$. The action $\#$ is not allowed. From q_{p_i}, the action i is not allowed while the other actions in Σ_n lead back the first state q_1^i and to the initial state q_0 both with probability $\frac{1}{2}$. Moreover, the action $\#$ leads back to the initial state (with probability 1). The disallowed actions lead to a bad state. We assume that the state spaces L_i of the A_i's are disjoint.

Game family $(P_n)_{n\in\mathbb{N}}$. The state space of P_n is the disjoint union of $Q_1, \ldots, Q_n$ and $\{q_0, \mathsf{Bad}\}$. The initial state is q_0, the final state is Bad. The probabilistic transition function is as follows:

- for all $1 \leq i \leq n$ and $\sigma \in \Sigma_n$, we have $\delta(q_0, \sigma,)(q_1^i) = \frac{1}{n}$;
- for all $1 \leq i \leq n$, $1 \leq j < p_i$, and $\sigma \in \Sigma_n$, $\sigma' \in \Sigma_n \setminus \{i\}$, we have $\delta(q_j^i, \sigma)(q_{j+1}^i) = \delta(q_j^i, \sigma)(q_0) = \delta(q_{p_i}^i, \sigma')(q_1^i) = \delta(q_{p_i}^i, \sigma')(q_1^i) = \frac{1}{2}$; and
- for all $1 \leq i \leq n$ and $1 \leq j < p_i$, we have $\delta(q_0, \#)(\mathsf{Bad}) = \delta(q_j^i, \#)(\mathsf{Bad}) = \delta(q_{p_i}^i, \#)(q_0) = 1$.

The initial state is q_0, there are two observations, the state $\{q_0\}$ is labelled by observation o_1, and the other states in $Q_1 \cup \cdots \cup Q_n$ that we refer to as the initial state and the loops respectively are labelled by observation o_2. Fig. 1 shows the game P_2.

Proof of Lemma 4. After the first transition from the initial state, player 1 has the following positive winning strategy. Let $p_n^* = \prod_{i=1}^n p_i$. While the POMDP is in the loops (assume that we have seen j times observation o_2 consecutively), if $1 \leq j < p_n^*$, then play any action i such that $j \mod p_i \neq 0$ (this is well defined since p_n^* is the lcm of $p_1, \ldots, p_n$), and otherwise play $\#$. It is easy to show that this strategy is winning for the safety condition, with probability 1.

For the second part of the result, assume towards contradiction that there exists a finite-memory randomized strategy $\hat{\alpha}$ that is positive winning for Player 1 and has less than p_n^* states (since p_n^* is exponential in $s_n^* = \sum_{i=1}^{n} p_i$, the result will follow). Let η be the least positive transition probability described by the finite-state strategy $\hat{\alpha}$. Consider any history of a play ρ that ends with o_1. We claim that the following properties hold: (a) with probability 1 either observation o_1 is visited again from ρ or the state Bad is reached; and (b) the state Bad is reached with a positive probability. The first property (property (a)) follows from the fact that for all actions the loops are left (the state q_0 or Bad is reached) with probability at least $\frac{1}{2}$. We now prove the second property by showing that the state Bad is reached with probability at least $\varDelta_n = \frac{1}{n} \cdot \frac{1}{(2 \cdot \eta)^{p_n^*}}$. To see this, consider the sequence of actions played by strategy $\hat{\alpha}$ after ρ when only o_2 is observed. Either $\#$ is never played, and then the action played by $\hat{\alpha}$ after a sequence of p_n^* states leads to Bad (the current state being then $q_{p_i}^i$ for some $1 \leq i \leq n$). This occurs with probability at least $\varDelta_n$; or $\#$ is eventually played, but since $\hat{\alpha}$ has less than p_n^* states, it has to be played after less than p_n^* steps, which also leads to Bad with probability at least $\varDelta_n$. The above two properties that (a) $o_1 \cup \{\mathsf{Bad}\}$ is reached with probability 1 from o_1, and (b) within p_n^* steps after a visit to o_1, the state Bad is reached with fixed positive probability, ensures that Bad is reached with probability 1. Hence $\hat{\alpha}$ is not positive winning. It follows that randomized strategies that are almost or positive winning in POMDPs with safety objective requires exponential memory.

5.2 Bounds for reachability objectives

We first present the bound for positive winning, and then for almost winning with reachability objectives in POMDPs.

Memory bounds. We now argue the memory bounds for pure and randomized strategies for positive winning with reachability objectives.

1. It follows from correctness argument of Theorem 1 that randomized memoryless strategies suffice for positive winning with reachability objectives in POMDPs.
2. We now argue that for pure strategies memory linear in the number of states is sufficient and necessary. The upper bound follows from the reduction to graph reachability. Given a POMDP G, consider the graph $\overline{G}$ constructed from G as in the correctness argument for Theorem 1. Given the starting state ℓ, if there is path in $\overline{G}$ to the target set T, then there is a path π of length at most $|L|$. The pure strategy for Player 1 in G can play the sequence of actions of the path π to ensure that the target observations $\mathcal{T}$ are reached with positive probability in G. The family of examples to show that pure strategies require linear memory can be constructed as follows: we construct a POMDP with deterministic transition function such that there is a unique path (sequence of actions) of length $O(|L|)$ to the target, and any deviation

leads to an absorbing state, and other than the target state every other state has the same observation. In this POMDP any pure strategy must remember the exact sequence of actions to be played and hence requires $O(|L|)$ memory.

It follows from the results of [1] that for almost winning with reachability objectives in POMDPs pure strategies with exponential memory suffices, and we now prove an exponential lower bound for randomized strategies.

Lemma 5. *There exists a family $(P_n)_{n\in\mathbb{N}}$ of* POMDP*s of size $O(p(n))$ for a polynomial p with a reachability objective such that the following assertions hold: (a) Player* 1 *has an almost winning strategy in each of these* POMDP*s; and (b) there exists a polynomial q such that every finite-memory randomized strategy for Player 1 that is almost winning in P_n has at least $2^{q(n)}$ states.*

Fix the action set as $\Sigma = \{\#, \mathsf{tick}\}$. The POMDP P'_n is composed of an initial state q_0 and n sub-MDPs H_i, each consisting of a loop over p_i states $q_1, \ldots, q_{p_i}$. From each state in the loops, the action tick can be played and leads to the next state in the loop (with probability 1). The action # can be played in the last state of each loop and leads to the Goal state. The objective is to reach Goal with probability 1. Actions that are not allowed lead to a sink state from which it is impossible to reach Goal. There is a unique observation that consists of the whole state space. Fig. 2 shows P'_2.

Proof of Lemma 5. First we show that Player 1 has an almost winning strategy in P'_k (from q_0). As there is only one observation, a strategy for Player 1 corresponds to a function $\alpha : \mathbb{N} \to \Sigma$. Consider the strategy α^* as follows: $\alpha^*(j) = \mathsf{tick}$ for all $0 \leq j < p^*_k$ and $\alpha^*(j) = \#$ for all $j \geq p^*_k$. It is easy to check that α^* ensures winning with certainty and hence almost winning.

For the second part of the result assume, towards a contradiction, that there exists a finite-memory randomized strategy $\hat{\alpha}$ that is almost winning and has less than p^*_k states. Clearly, $\hat{\alpha}$ cannot play # before the $(p^*_k + 1)$-th round since in one of the subMDPs H_i would not be in $q^i_{p_i}$ and therefore Player 1 would loose with probability at least $\frac{1}{n}$. Note that the state reached by the strategy automaton defining $\hat{\alpha}$ after p^*_k rounds has necessarily been visited in a previous round. Since β has to play # eventually to reach Goal, this means that # must have been played in some round $j < p^*_k$, when at least one of the subgames subgames H_i was not in location $q^i_{p_i}$, so that Player 1 would have already lost with probability at least $\frac{1}{n} \cdot \eta$, where η is the least positive probability specified by $\hat{\alpha}$. This is in contradiction with our assumption that $\hat{\alpha}$ is an almost winning strategy.

Bounds for Büchi and coBüchi objectives. An exponential upper bound for memory of pure strategies for almost winning of Büchi objectives follows from the results of [1], and the matching lower bound for randomized strategies follows from our result for reachability objectives. Since positive winning is undecidable for Büchi objectives there is no bound on memory for pure or randomized strategies for positive winning. An exponential upper bound for memory of pure strategies for positive winning of coBüchi objectives follows from the correctness proof of Theorem 3 that iteratively combines the positive winning strategies for safety and reachability to obtain a positive winning strategy for coBüchi objective. The matching lower bound for randomized strategies follows from our result for safety objectives. Since almost winning is undecidable for coBüchi objectives there is no bound on memory for pure or randomized strategies for positive winning. This gives us the following theorem (also summarized in Table 2), which is in contrast to the results for MDPs with perfect observation where pure memoryless strategies suffices for almost and positive winning for all parity objectives.

Theorem 6. *The optimal memory bounds for strategies in* POMDP*s is as follows.*

1. *Reachability objectives: for positive winning randomized memoryless strategies exist, and linear memory is necessary and sufficient for pure strategies; and for almost winning exponential memory is necessary and sufficient for both pure and randomized strategies.*
2. *Safety objectives: for positive winning and almost winning exponential memory is necessary and sufficient for both pure and randomized strategies.*
3. *Büchi objectives: for almost winning exponential memory is necessary and sufficient for both pure and randomized strategies; and there is no bound on memory for pure and randomized strategies for positive winning.*
4. *coBüchi objectives: for positive winning exponential memory is necessary and sufficient for both pure and randomized strategies; and there is no bound on memory for pure and randomized strategies for almost winning.*

	Pure Positive	Randomized Positive	Pure Almost	Randomized Almost
Reachability	Linear	Memoryless	Exponential	Exponential
Safety	Exponential	Exponential	Exponential	Exponential
Büchi	No Bound	No Bound	Exponential	Exponential
coBüchi	Exponential	Exponential	No Bound	No Bound
Parity	No Bound	No Bound	No Bound	No Bound

Table 2. Optimal memory bounds for pure and randomized strategies for positive and almost winning.

References

1. C. Baier, N. Bertrand, and M. Größer. On decision problems for probabilistic Büchi automata. In *FoSSaCS: Foundations of Software Science and Computational Structures*, LNCS 4962, pages 287–301. Springer, 2008.

2. D. Berwanger and L. Doyen. On the power of imperfect information. In *FSTTCS 2008*, Dagstuhl Seminar Proceedings 08004. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), 2008.
3. A. Bianco and L. de Alfaro. Model checking of probabilistic and nondeterministic systems. In *FSTTCS 95: Software Technology and Theoretical Computer Science*, volume 1026 of *Lecture Notes in Computer Science*, pages 499–513. Springer-Verlag, 1995.
4. K. Chatterjee, L. Doyen, T. A. Henzinger, and J.-F. Raskin. Algorithms for omega-regular games of incomplete information. *Logical Methods in Computer Science*, 3(3:4), 2007.
5. K. Chatterjee, M. Jurdziński, and T.A. Henzinger. Quantitative stochastic parity games. In *SODA 04: ACM-SIAM Symposium on Discrete Algorithms*, pages 114–123, 2004. Technical Report: UCB/CSD-3-1280 (October 2003).
6. L. de Alfaro. *Formal Verification of Probabilistic Systems.* PhD thesis, Stanford University, 1997. Technical Report STAN-CS-TR-98-1601.
7. H. Gimbert. Personal communication.
8. A. Kechris. *Classical Descriptive Set Theory.* Springer, 1995.
9. M.L. Littman. *Algorithms for sequential decision making.* PhD thesis, Brown University, 1996.
10. C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12:441–450, 1987.
11. R. Segala. *Modeling and Verification of Randomized Distributed Real-Time Systems.* PhD thesis, MIT, 1995. Technical Report MIT/LCS/TR-676.
12. W. Thomas. Languages, automata, and logic. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 3, Beyond Words, chapter 7, pages 389–455. Springer, 1997.
13. M.Y. Vardi. Automatic verification of probabilistic concurrent finite-state systems. In *Proceedings of the 26th Annual Symposium on Foundations of Computer Science*, pages 327–338. IEEE Computer Society Press, 1985.