Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Finite-Memory Strategies in POMDPs with Long-Run Average Objectives

Krishnendu Chatterjee, Raimundo Saona, Bruno Ziliotto

informs.
http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 47, No. 1, February 2022, pp. 100–119
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Finite-Memory Strategies in POMDPs with Long-Run Average Objectives

Krishnendu Chatterjee,[a] Raimundo Saona,[a] Bruno Ziliotto[b]

[a] Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria; [b] Centre de Recherche en Mathématiques de la Décision, Centre National de la Recherche Scientifique, Université Paris Dauphine, Université PSL, 75016 Paris, France

Contact: krishnendu.chatterjee@ist.ac.at, https://orcid.org/0000-0002-4561-241X (KC); raimundo.saona@ug.uchile.cl, https://orcid.org/0000-0001-5103-038X (RS); bruno.ziliotto01@gmail.com, http://orcid.org/0000-0002-4448-1411 (BZ)

Received: February 14, 2020
Revised: July 16, 2020
Accepted: August 8, 2020
Published Online in Articles in Advance:
April 6, 2021

MSC2020 Subject Classification: 90C39, 90C40, 37A50, 03D15

https://doi.org/10.1287/moor.2020.1116

Abstract. Partially observable Markov decision processes (POMDPs) are standard models for dynamic systems with probabilistic and nondeterministic behaviour in uncertain environments. We prove that in POMDPs with long-run average objective, the decision maker has approximately optimal strategies with finite memory. This implies notably that approximating the long-run value is recursively enumerable, as well as a weak continuity property of the value with respect to the transition function.

Funding: This work was supported by the European Cooperation in Science and Technology [Project GAMENET], the Austrian Science Fund [NFN Grant RiSE/SHiNE S11407], the Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) of Chile [Grant PII 20150140], ECOS-CONICYT [Grant C15E03], ERC CoG 863818 (ForM-SMArt), and the Fondation Mathématique Jacques Hadamard [Program PGMO RSG 2018-0031H] with the support of EDF, Thales, Orange, and Criteo.

Keywords: finite state • Markov • dynamic programming • computational complexity • analysis of algorithms

1. Introduction

In a partially observable Markov decision process (POMDP), at each stage, the decision maker chooses an action that determines, together with the current state, a stage reward and the distribution over the next state. The state dynamic is imperfectly observed by the decision maker, who receives a stage signal on the current state before playing. Thus, POMDPs generalize the Markov decision process (MDP) model of Bellman [3].

POMDPs are widely used in prominent applications such as in computational biology (Durbin et al. [10]), software verification (Cerný et al. [8]), and reinforcement learning (Kaelbling et al. [16]), to name a few. Even special cases of POMDPs, namely, probabilistic automata or blind MDPs, where there is only one signal, are also standard models in several applications (Bukharaev [7], Paz [19], Rabin [20]).

In many of these applications, the duration of the problem is huge. Thus, considerable attention has been devoted to the study of POMDPs with long durations. A standard way is to consider the long-run objective criterion, where the total reward is the expectation of the inferior limit average reward (see Arapostathis et al. [1] for a survey). The value for this problem is known to coincide with several classical definitions of long-run values (asymptotic value, uniform value, general uniform value, long-run average value, and uncertain-duration process value; Neyman and Sorin [18], Renault [21], Renault and Venel [22], Rosenberg et al. [23], Venel and Ziliotto [27]) and has been characterized in Renault and Venel [22]. In this paper, we will simply name this common object *value*. Thus, strong results are available concerning the existence and characterization of the value.

This is in sharp contrast with the study of long-run optimal strategies. Indeed, before our work, little was known about the sophistication of strategies that approximate the value. It has been shown that (i) stationary strategies approximate the value in MDPs (Blackwell [4]), and (ii) belief-stationary strategies approximate the value in blind MDPs (Rosenberg et al. [23]) and POMDPs with an ergodic structure (Borkar [6]).

Our main contributions are as follows:

- *Strategy complexity.* We show that for every POMDP with long-run average objectives, for every $\varepsilon > 0$, there is a finite-memory strategy (i.e., generated by a finite state automaton) that achieves expected reward within ε of the optimal value. In the case of blind MDPs, finite memory is equivalent to finite recall (i.e., decisions are defined using only the last actions), but finite recall cannot achieve ε-approximations in general POMDPs.
- *Computational complexity.* An important consequence of our above result is that the decision version of the approximation problem for POMDPs with long-run average objectives (see Definition 3.1) is *recursively enumerable* (*r.e.*) but not decidable. Our results on strategy complexity imply the recursively enumerable upper bound, and the lower bound is a consequence of Madani et al. [17].

- *Value property.* The long-run reward of a finite-memory strategy is robust upon small perturbations of the transition function, where the notion of perturbation over the transition function is defined as in Solan [25] and Solan and Vieille [26]. This implies lower semicontinuity of the value function upon such small perturbations.

A natural question would be to ask for an upper bound on the size of the memory needed to generate ε-optimal strategies, in terms of the data of the POMDP. In fact, a previous *undecidability* result (Madani et al. [17]) shows that such an upper bound cannot exist (see Subsection 3.1). Thus, the existence of ε-optimal strategies with finite memory is, in some sense, the best possible result one can have in terms of strategy complexity.

2. Model and Statement of Results

2.1. Model

Throughout this paper, we mostly use the following notation: (i) sets are denoted by calligraphic letters, for example, $\mathcal{A}, \mathcal{H}, \mathcal{K}$, and $\mathcal{S}$; (ii) elements of these sets are denoted by lowercase letters, for example, a, h, k, and s; and (iii) random elements with values in these sets are denoted by uppercase letters, for example, A, H, K, and S. For a set $\mathcal{C}$, denote by $\Delta(\mathcal{C})$ the set of probability measure distributions over $\mathcal{C}$, and by δ_c the Dirac measure at some element $c \in \mathcal{C}$. We will slightly abuse notation by not making a distinction between a probability measure (which can be evaluated on events) and its corresponding probability density (which can be evaluated on elements).

Consider a POMDP $\Gamma = (\mathcal{K}, \mathcal{A}, \mathcal{S}, q, g)$, with finite state space $\mathcal{K}$, finite action set $\mathcal{A}$, finite signal set $\mathcal{S}$, transition function $q : \mathcal{K} \times \mathcal{A} \to \Delta(\mathcal{K} \times \mathcal{S})$, and reward function $g : \mathcal{K} \times \mathcal{A} \to [0, 1]$.

Given $p_1 \in \Delta(\mathcal{K})$, called the *initial belief*, the POMDP starting from p_1 is denoted by $\Gamma(p_1)$ and proceeds as follows:

- An initial state K_1 is drawn from p_1. The decision maker knows p_1 but does not know K_1.
- At each stage $m \geq 1$, the decision maker takes some action $A_m \in \mathcal{A}$. This action determines a stage reward $G_m := g(K_m, A_m)$, where K_m is the (random) state at stage m. Then, the pair (K_{m+1}, S_m) is drawn from $q(K_m, A_m)$. The next state is K_{m+1} and the decision maker is informed of the signal S_m, but neither the reward G_m nor the state K_{m+1}.

At stage m, the decision maker remembers all the past actions and signals, which is called the *history before stage* m. Let $\mathcal{H}_m := (\mathcal{A} \times \mathcal{S})^{m-1}$ be the set of histories before stage m, with the convenient notation $(\mathcal{A} \times \mathcal{S})^0 := \{\emptyset\}$. A strategy is a mapping $\sigma : \cup_{m \geq 1} \mathcal{H}_m \to \mathcal{A}$. The set of strategies is denoted by Σ. The randomness introduced by the transition function, $q : \mathcal{K} \times \mathcal{A} \to \Delta(\mathcal{K} \times \mathcal{S})$, suggests that a history $h_m \in \mathcal{H}_m$ can occur under many sequences of states $(k_1, k_2, \ldots, k_{m-1})$. The infinite sequence $(k_1, a_1, s_1, k_2, a_2, s_2, \ldots)$ is called a *play*, and the set of all plays is denoted by Ω.

For $p_1 \in \Delta(\mathcal{K})$ and $\sigma \in \Sigma$, define $\mathbb{P}^{p_1}_\sigma$ as the law induced by σ and the initial belief p_1 on the set of plays of the game $\Omega = (\mathcal{K} \times \mathcal{A} \times \mathcal{S})^{\mathbb{N}}$, and $\mathbb{E}^{p_1}_\sigma$ as the expectation with respect to this law. For simplicity, identify $\mathcal{K}$ with the set of extremal points of $\Delta(\mathcal{K})$.

Let

$$\gamma^{p_1}_\infty(\sigma) := \mathbb{E}^{p_1}_\sigma\left(\liminf_{n \to +\infty} \frac{1}{n} \sum_{m=1}^{n} G_m\right),$$

and

$$v_\infty(p_1) := \sup_{\sigma \in \Sigma} \gamma^{p_1}_\infty(\sigma).$$

The term $\gamma^{p_1}_\infty(\sigma)$ is the long-term reward given by strategy σ, and $v_\infty(p_1)$ is the optimal long-term reward, called *value*, defined as the supremum long-term reward over all strategies.

Remark 2.1. It has been shown that v_∞ coincides with the limit of the value of the n-stage problem and λ-discounted problem, as well as the uniform value and weighted uniform value (see Renault [21], Renault and Venel [22], Rosenberg et al. [23], Venel and Ziliotto [27]). In particular, we have

$$\begin{aligned} v_\infty(p_1) &= \lim_{n \to +\infty} \sup_{\sigma \in \Sigma} \mathbb{E}^{p_1}_\sigma\left(\frac{1}{n} \sum_{m=1}^{n} G_m\right) \\ &= \lim_{\lambda \to 0} \sup_{\sigma \in \Sigma} \mathbb{E}^{p_1}_\sigma\left(\sum_{m \geq 1} \lambda(1-\lambda)^{m-1} G_m\right) \\ &= \sup_{\sigma \in \Sigma} \liminf_{n \to +\infty} \mathbb{E}^{p_1}_\sigma\left(\frac{1}{n} \sum_{m=1}^{n} G_m\right). \end{aligned}$$

Remark 2.2. In the literature, the concept of strategy that we defined is often called *pure strategy*, by contrast with *behaviour strategies* that use randomness by allowing strategies of the form $\sigma\colon \cup_{m\geq 1} \mathcal{H}_m \to \Delta(\mathcal{A})$. By Kuhn's theorem, enlarging the set of pure strategies to behaviour strategies does not change v_∞ (see Feinberg [11], Venel and Ziliotto [27]]), and thus does not change our results.

Definition 2.3 (Blind MDP)**.** A POMDP is called a *blind MDP* if the signal set is a singleton.

Note that in a blind MDP, signals do not convey any relevant information. Therefore, a strategy is simply an infinite sequence of actions $(a_1, a_2, \ldots) \in \mathcal{A}^{\mathbb{N}}$.

2.2. Contribution

We start by defining several classes of strategies. Recall that $\Gamma(p_1)$ is the POMDP Γ starting from p_1, which is known to the player.

Definition 2.4 (ε-Optimal Strategy)**.** Let $p_1 \in \Delta(\mathcal{K})$ and $\varepsilon > 0$. A strategy $\sigma \in \Sigma$ is *ε-optimal* in $\Gamma(p_1)$ if

$$\gamma_\infty^{p_1}(\sigma) \geq v_\infty(p_1) - \varepsilon.$$

Definition 2.5 (Finite-Memory Strategy)**.** A strategy σ is said to have *finite memory* if it can be modeled by a finite-state transducer. Formally, $\sigma = (\sigma_u, \sigma_a, \mathcal{M}, m_0)$, where $\mathcal{M}$ is a finite set of memory states, m_0 is the initial memory state, $\sigma_a : \mathcal{M} \to \mathcal{A}$ is the action selection function, and $\sigma_u{:}\mathcal{M} \times \mathcal{A} \times \mathcal{S} \to \mathcal{M}$ is the memory update function.

Definition 2.6 (Finite-Recall Strategy)**.** A strategy σ is said to have *finite recall* if there exists a constant $M > 0$ such that for all $h_M \in \mathcal{H}_M$, and for all $m > M$ and $h_{m-M} \in \mathcal{H}_{m-M}$, we have that $\sigma(h_{m-M}, h_M)$ does not depend on h_{m-M}.

Remark 2.7. For blind MDPs, finite-recall and finite-memory strategies coincide with the set of *eventually periodic strategies*: a strategy $\sigma = (a_1, a_2, \ldots)$ is eventually periodic if there exists $T \geq 1$ and $N \geq 1$ such that for all $m \geq N$, $a_{m+T} = a_m$. This property does not extend to general POMDPs (see Proposition 2.12): any finite-recall strategy has finite memory, but the inverse is not true.

Remark 2.8. Finite-memory strategies and finite-recall strategies have been investigated in the Shapley zero-sum stochastic game model (Shapley [24]). In this framework, none of these strategies is enough to approximate the value, and a long-standing open problem is whether finite-memory strategies *with a clock* are good enough (see Hansen et al. [13, 14] for more details on this topic).

Our main result is the following theorem.

Theorem 2.9. *For every POMDP Γ, initial belief p_1, and $\varepsilon > 0$, there exists an ε-optimal finite-memory strategy in $\Gamma(p_1)$.*

Remark 2.10. A previous complexity result (Madani et al. [17]) shows that the size of the memory cannot be bounded from above in terms of the data of the POMDP (see Subsection 3.1).

Corollary 2.11. *For every blind MDP Γ, initial belief p_1, and $\varepsilon > 0$, there exists an ε-optimal finite-memory strategy in $\Gamma(p_1)$, and thus the strategy is eventually periodic and has finite recall.*

Last, finite recall is not enough to ensure ε-optimality in general POMDPs.

Proposition 2.12. *There exists a POMDP, and $\varepsilon > 0$, with no ε-optimal finite-recall strategy.*

The rest of this paper is organized as follows. Section 3 explains the consequences of our result in terms of complexity and model robustness. Section 4 introduces examples used to prove negative results and to illustrate our techniques. Section 5 introduces two key lemmas and shows that they imply Theorem 2.9. Section 6 proves one of the two lemmas and develops what we call *supersupport*-based strategies in detail. Missing proofs are in the appendices.

3. Consequences of the Results

3.1. Complexity

3.1.1. Decidability. A decision problem consists in deciding between two options given an input (accepting or rejecting) and its complexity is characterized by Turing machines. A Turing machine takes an input and, if it halts, it either accepts or rejects it. If it halts for all possible inputs in a finite number of steps, then the Turing machine is considered an algorithm. An algorithm solves a decision problem if it takes the correct decision for all inputs. The class of decision problems that are solvable by an algorithm is called *decidable*. Two natural generalizations of decidable problems are *recursively enumerable* and *corecursively enumerable* (*cor.e.*). The r.e. (respectively, cor.e.) decision problems are those for which there is a Turing machine that accepts (respectively, rejects) every input that should be accepted (respectively, rejected) according to the problem, but, on other inputs, it needs not halt.

Notice that the class of decidable problems is the intersection of r.e. and cor.e. In this work, the algorithmic problem of interest is the following.

Definition 3.1 (Decision Version of Approximating the Value)**.** Let $p_1 \in \Delta(\mathcal{K})$. Given $x \in [0,1]$, $\varepsilon > 0$ such that $v_\infty(p_1) > x + \varepsilon$ or $v_\infty(p_1) < x - \varepsilon$, the problem consists in deciding which one is the case: to accept means to prove that $v_\infty(p_1) > x + \varepsilon$ holds, whereas to reject means to prove the opposite.

3.1.2. Previous Results and Implication of Our Result.

It is known that the decision version of the approximation problem is not decidable (Madani et al. [17]; even for blind MDPs). However, the complexity characterization has been open. Thanks to Theorem 2.9, we can design a Turing machine that accepts every input that should be accepted for this problem.

Consider playing a finite-memory strategy σ. Then, the dynamics of the game can be described by a finite Markov chain. Therefore, the reward obtained by playing σ (i.e., $\gamma_\infty^{p_1}(\sigma)$) can be deduced from its stationary measure, which can be computed in polynomial time by solving a linear programming problem (Filar and Vrieze [12, section 2.9, p. 70]). Our protocol checks the reward given by every finite-memory strategy to approximate the value of the game $v_\infty(p_1)$. By Theorem 2.9, if $v_\infty(p_1) > x + \varepsilon$ holds, a finite-memory strategy that achieves a reward strictly greater than $(x + \varepsilon)$ will be eventually found and our protocol will accept the input. On the other hand, if $v_\infty(p_1) < x - \varepsilon$, the protocol will never find out that this is the case because there are infinitely many finite-memory strategies, so it will not halt. Thus, our result establishes that the approximation version of the problem is r.e., and the previously known results imply that the problem is not decidable. Formally, we have the following result.

Corollary 3.2. *The decision version of approximating the value is r.e. but not decidable.*

Remark 3.3. The former paragraph shows that no upper bound on the size of the memory used by ε-optimal strategies can be proved. Indeed, if such a bound existed, one could modify the previous algorithm in the following way: reject the input if every finite-memory strategy of size lower than the bound has been enumerated. This would imply that the decision version of approximating the value is decidable, which is a contradiction.

3.2. Objective Comparison

In this section, we contrast our results with other natural objectives.

Recall that the value of $\Gamma(p_1)$ is defined as

$$v_\infty(p_1) = \sup_{\sigma \in \Sigma} \mathbb{E}_\sigma^{p_1}\left(\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} G_m\right).$$

We say this is the *liminf-average* objective. Consider replacing $\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} G_m$ by (i) $\limsup_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} G_m$, which we call the *limsup-average* objective, or (ii) $\limsup_{n\to\infty} G_n$, which we call the *limsup* objective.

Proposition 3.4. *For both limsup-average and limsup objectives, there exists a POMDP, and $\varepsilon > 0$, with no ε-optimal finite-memory strategy.*

This negative result, proved in Section 4.1.2, does not imply any computational complexity characterization for the limsup-average or limsup objective, and whether the approximation of the value problem for limsup-average objectives is recursively enumerable remains open. However, it shows that any approach based on finite-memory strategies cannot establish recursively enumerable bounds for the approximation problem.

Let us focus on the limsup objective. The limsup objective is arguably simpler than the liminf-average objective, and, to formalize this statement, we can compare the complexity of the objects themselves irrespective of any particular context or model (such as POMDPs). The *Borel hierarchy* describes the complexity of an objective by the number of quantifier alternations needed to describe it. Its construction is similar to that of the Borel σ-algebra, or σ-field, and is defined as follows.

Definition 3.5 (Borel Hierarchy)**.** Consider $h_m \in \mathcal{H}_m = (\mathcal{A} \times \mathcal{S})^{m-1}$ a finite history of the game. The cylinder set generated by h_m is given by $\{h_m\} \times (\mathcal{A} \times \mathcal{S})^{\mathbb{N}}$. Finite intersections, unions, and complements of the cylinder sets generated by finite histories form the first level in the hierarchy. Countable unions of the first level form Σ_1, and countable intersections form Π_1. The next level is always obtained from the previous one: countable unions of Π_i give Σ_{i+1}, and countable intersections of Σ_i give Π_{i+1}. The nested sequence of families of problems $\{\Sigma_i \cup \Pi_i\}_{i \geq 1}$ is called a Borel hierarchy.

For example, the limsup objective can be described as a countable intersection of countable unions of rewards: given a family of sets $(\mathcal{C}_n)_{n\geq 1}$, $\limsup_{n\to\infty} \mathcal{C}_n = \cap_{n\geq 1} \cup_{m\geq n} \mathcal{C}_m$. The formal result is the following (see Chatterjee [9]).

Proposition 3.6. *The limsup objective is Π_2-complete, that is, complete for the second level of the Borel hierarchy, whereas the liminf-average objective is Π_3-complete, that is, complete for the third level of the Borel hierarchy.*

Whereas the notion of Borel hierarchy characterizes the topological complexity for objectives, a similar notion of *arithmetic hierarchy* characterizes the computational complexity for decision problems.

Definition 3.7 (Arithmetic Hierarchy)**.** Denote by Σ_0^1 the class of r.e. problems and by Π_0^1 the cor.e. problems. For $i > 1$, define Σ_0^i as the class of problems solved by Turing machines with access to oracles for Π_0^{i-1}, and Π_0^i is similarly defined with oracles for Σ_0^{i-1}. The nested sequence of families of problems $\{\Sigma_0^i \cup \Pi_0^i\}_{i\geq 1}$ is called an arithmetic hierarchy.

By Corollary 3.2, we have that a POMDP with liminf-average objective is in $\Sigma_0^1 \setminus (\Sigma_0^1 \cap \Pi_0^1)$. On the other hand, it was shown in Baier et al. [2] and Bonet and Geffner [5] that a POMDP with a limsup objective with Boolean rewards is Σ_0^2-complete.

We conclude this section with a summary chart contrasting the liminf-average and limsup objectives; see Figure 1. The surprising result is the complexity switch: the limsup objective has lower Borel hierarchy complexity but higher arithmetic hierarchy complexity in the context of POMDPs.

3.3. Robust ε-Optimal Strategies

Consider a POMDP $\Gamma = (\mathcal{K}, \mathcal{A}, \mathcal{S}, q, g)$. It is well known that the value function is continuous with respect to perturbations of the reward function g and the initial belief p_1. Now, we show a robustness result concerning the transition function q.

In applications, just as in any stochastic model, the structure of the model is decided first, and then the specific probabilities are either estimated or fixed. The values of transition probabilities are approximations: an ε-perturbation of these probabilities is not expected to have an impact on the modelling. In our setting, the transitions are encoded in the function $q : \mathcal{K} \times \mathcal{A} \to \Delta(\mathcal{K} \times \mathcal{S})$, and we would expect some robustness against perturbations of the values it takes.

The notion of perturbation over q is measured as in Solan [25] and Solan and Vieille [26], where perturbations in each transition probability are measured as relative differences, not additive differences. Formally, define the semimetric

$$d(q, q') = \max_{\substack{k\in\mathcal{K},\, a\in\mathcal{A} \\ k'\in\mathcal{K},\, s\in\mathcal{S}}} \left\{ \frac{q(k,a)(k',s)}{q'(k,a)(k',s)}, \frac{q'(k,a)(k',s)}{q(k,a)(k',s)} \right\} - 1 \,.$$

Under this notion, and taking q and q' close to each other, we can prove the existence of strategies that are approximately optimal for the POMDP corresponding to q and perform almost as well when they are applied to the POMDP corresponding to q'. To formally state this notion of robustness, let us give the following definition.

Definition 3.8 (Robust Strategies)**.** Given a POMDP Γ with transition function q and an initial belief $p_1 \in \Delta(\mathcal{K})$, we say that σ is a *robust strategy* for $\Gamma(p_1)$ if the following condition holds: $\forall \eta > 0\ \exists \delta > 0$ such that

$$d(q, q') \leq \delta \quad \Rightarrow \quad \gamma_\infty'^{p_1}(\sigma) \geq \gamma_\infty^{p_1}(\sigma) - \eta \,,$$

Figure 1. Objective comparison in POMDPs.

Objective comparison in POMDPs		
Objective	Borel hierarchy	Arithmetic Hierarchy
limsup	Π_2-complete	Σ_0^2-complete
liminf-average	Π_3-complete	$\Sigma_0^1 \setminus (\Sigma_0^1 \cap \Pi_0^1)$

where γ_∞ is the long-term reward in Γ, and γ'_∞ is the long-term reward in $\Gamma' = (\mathcal{K}, \mathcal{A}, \mathcal{S}, g, q')$.

Lemma 3.9. *Any finite-memory strategy is robust. Thus, in any POMDP and for any $\varepsilon > 0$, there exists a robust ε-optimal finite-memory strategy.*

Proof. Let $\Gamma = (\mathcal{K}, \mathcal{A}, \mathcal{S}, g, q)$ be a POMDP. Consider $\sigma = (\sigma_u, \sigma_a, \mathcal{M}, m_0)$ a finite-memory strategy for $\Gamma(p_1)$. Playing σ from p_1 induces a Markov chain $(Y_n)_{n\geq 1}$ on $\mathcal{K} \times \mathcal{S} \times \mathcal{M}$. Define $\tilde{g} : \mathcal{K} \times \mathcal{M} \to [0,1]$ by $\tilde{g}(k, m) := g(k, \sigma_a(m))$.

Now, consider the 0-player stochastic game with reward $\tilde{g}$ and transitions given by the kernel of the Markov chain $(Y_n)_{n\geq 1}$. Let $s_0 \in \mathcal{S}$ be any signal. By definition, for any $k \in \mathcal{K}$, the value of this stochastic game starting from (k, s_0, m_0) coincides with $\gamma^k_\infty(\sigma)$. Using Solan [25, theorem 6, p. 841], we deduce that

$$\gamma'^k_\infty(\sigma) \geq \gamma^k_\infty(\sigma) - 4|\mathcal{K}||\mathcal{S}||\mathcal{M}| d(q, q') \,.$$

Integrating k over p_1 yields

$$\gamma'^{p_1}_\infty(\sigma) \geq \gamma^{p_1}_\infty(\sigma) - 4|\mathcal{K}||\mathcal{S}||\mathcal{M}| d(q, q') \,.$$

Taking $\delta = \eta(4|\mathcal{K}||\mathcal{S}||\mathcal{M}|)^{-1}$, we conclude that σ is a robust strategy. By Theorem 2.9, for all $\varepsilon > 0$, there exists an ε-optimal finite-memory strategy, which is thus robust. □

Corollary 3.10. *Let $\mathcal{K}, \mathcal{A}$, and $\mathcal{S}$ be finite sets; $g : \mathcal{K} \times \mathcal{A} \to \mathbb{R}$ a reward function; and $p_1 \in \Delta(\mathcal{K})$ an initial belief. The mapping from $(\Delta(\mathcal{K} \times \mathcal{S})^{\mathcal{K}\times\mathcal{A}}, d)$ to $\mathbb{R}$ that maps each transition function q to the value at p_1 of the POMDP $(\mathcal{K}, \mathcal{A}, \mathcal{S}, g, q)$ is lower semicontinuous.*

Proof. Let $q \in \Delta(\mathcal{K} \times \mathcal{S})^{\mathcal{K}\times\mathcal{A}}$. Let $\Gamma = (\mathcal{K}, \mathcal{A}, \mathcal{S}, q, g)$. By the previous lemma, for all $\varepsilon > 0$, there exists σ_ε, a robust ε-optimal strategy in $\Gamma(p_1)$. Take $\eta = \varepsilon$. By robustness of σ_ε, there exists $\delta > 0$ such that for all $q' \in \Delta(\mathcal{K} \times \mathcal{S})^{\mathcal{K}\times\mathcal{A}}$, we have that if $d(q, q') \leq \delta$, then $\gamma'^{p_1}_\infty(\sigma_\varepsilon) \geq \gamma^{p_1}_\infty(\sigma_\varepsilon) - \varepsilon$. Also, by ε-optimality of σ_ε, we have that $\gamma^{p_1}_\infty(\sigma_\varepsilon) \geq v_\infty(p_1) - \varepsilon$. Then,

$$v'_\infty(p_1) \geq \gamma'^{p_1}_\infty(\sigma_\varepsilon) \geq v_\infty(p_1) - 2\varepsilon \,.$$

Taking $\varepsilon \to 0$, we conclude that

$$\liminf_{q' \to q} v'_\infty(p_1) \geq v_\infty(p_1) \,,$$

and thus v_∞ is lower semicontinuous with respect to q. □

4. Examples

In this section, we introduce examples to prove negative results (Propositions 2.12 and 3.4) and to illustrate our techniques later on.

4.1. Negative Results

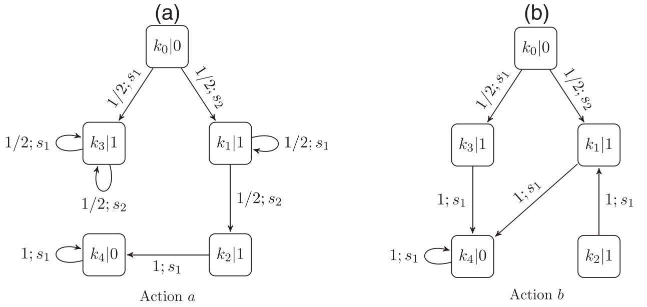
Let us prove Propositions 2.12 and 3.4 by presenting an example for each statement.

4.1.1. Proof of Proposition 2.12. We will prove that there exists a POMDP and $\varepsilon > 0$ with no ε-optimal finite-recall strategy by an explicit construction. Recall that a strategy has finite recall if it uses only a finite number of the last stages in the current history to decide the next action (see Definition 2.6). Therefore, our construction should have the property that, for any finite-recall strategy, there is a pair of finite histories such that

1. the last stages are identical, that is, the player did the same actions and received the same signals in the last part of both histories (but the starting point was different),
2. taking the same decision in both histories leads to losing some reward that cannot be compensated in the long run, and
3. the previous loss does not decrease to zero by increasing the amount of memory.

Example 4.1. Consider the POMDP $\Gamma = (\mathcal{K}, \mathcal{A}, \mathcal{S}, q, g)$ with five states: $k_0, k_1, \ldots, k_4$. The initial state is k_0 and players know it (formally, the initial belief is δ_{k_0}). The state k_4 is an absorbing state from where it is impossible to get out and rewards are zero. The states k_1 and k_2 form a subgame where the optimal strategy is trivial. This is the same for the state k_3. From k_0, a random initial signal is given indicating which subgame the state moved to. The key idea is that

Figure 2. Finite recall is not enough for POMDPs.



there is an arbitrarily long sequence of actions and signals that can be gotten in both subgames, but the optimal strategy behaves differently in each of them. Therefore, to forget the initial signal of the POMDP leads to at most half of the optimal value.

Figure 2 is a representation of Γ, first under action a and then action b. Each state is followed by the corresponding reward, and each arrow includes the probability for the corresponding transition along with the signal obtained.

The subgame of k_1 and k_2 has a unique optimal strategy: play action a until receiving signal s_2, then play action b once and repeat. The value of this subgame is 1, and deviating from the prescribed strategy would lead to a long-run reward of 0. Similarly, the value of the subgame of k_3 has a unique optimal strategy: to always play action a. Again, the value of this subgame is 1, and playing any other strategy would lead to a long-run reward of 0.

By the previous discussion, the value of this game starting from k_0 is 1. On the other hand, the maximum value obtained by strategies with finite recall is only $1/2$, by playing, for example, always action a. Finite-recall strategies achieve at most $1/2$ because, no matter how much finite recall there is, by playing the game, the decision maker faces a history of having played always action a and always receiving a signal s_1, except for the last signal, which is s_2. Then, if action b is played, the second subgame is lost; if action a is played, the first subgame is lost. That is why, for any $0 < \varepsilon < 1/2$, there is no ε-optimal finite-recall strategy for this POMDP.

4.1.2. Proof of Proposition 3.4. We will show that for the limsup-average and limsup objectives there is a blind MDP where there is no ε-optimal finite-memory strategy. For both cases, the example is constructed with the following idea in mind. To achieve the optimal value, the decision maker needs to play an action a_1 for some period, then play another action a_2 and repeat the process. The key is to require that the length of the period gets longer as the game progresses. This kind of strategy cannot be achieved with finite-memory strategies.

For the limsup-average objective, the blind MDP example is due to Venel and Ziliotto [28] and is presented below.

Example 4.2. Consider two states k_0 and k_1, and the player receives a reward only when the state is k_1. To reach k_1, the decision maker can play action *change* and move between the two states. By playing action *wait*, the state does not change.

Figure 3 is a representation of the game.

Consider the initial belief $p_1 = \frac{1}{2} \cdot \delta_{k_0} + \frac{1}{2} \cdot \delta_{k_1}$, the uniform distribution. It is easy to see that finite-memory strategies (or, equivalently, finite-recall strategies) cannot achieve more than $1/2$. On the other hand, the value of this game with the limsup-average objective is 1 and is guaranteed by the following strategy:

$$\sigma = (wait)^{2^{0^2}} (change)(wait)^{2^{1^2}} \cdots (change)(wait)^{2^{N^2}} \cdots .$$

Figure 3. Finite recall is not enough for POMDPs.

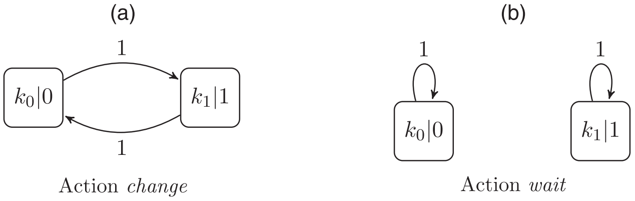
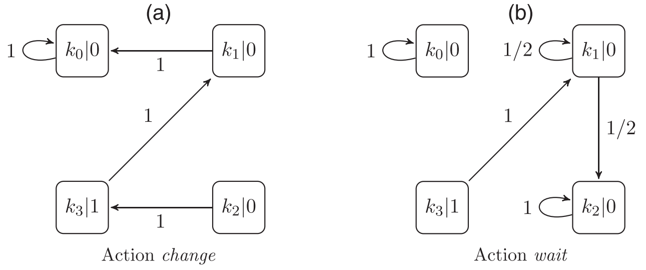


Figure 4. Finite memory is not enough for the limsup objective.



Hence, finite-memory strategies do not guarantee any approximation for POMDPs with the limsup-average objective.

For the limsup objective, the blind MDP example is the following.

Example 4.3. Consider four states (k_0, k_1, k_2, and k_3) and two actions: *wait* (w) and *change* (c). The initial state is k_1 and players know it. In k_1, if w is played, then the state moves to k_2 with probability $1/2$ and stays with probability $1/2$; if c is played, then the absorbing state k_0 is reached. From state k_2, if we play w, we stay in the same state; if we play c, we move to state k_3. From k_3, the only state that has a positive reward, if we play any action, we return to the initial state k_1.

Figure 4 is a representation of the game.

In this blind MDP (see Baier et al. [2], Bonet and Geffner [5]), for the limsup objective, for any $\varepsilon > 0$, there is an infinite-memory strategy that guarantees $1 - \varepsilon$, so the value of the game is 1. On the other hand, applying any finite-memory strategy (or, equivalently, finite-recall strategies) yields a limsup reward of 0. Hence, finite-memory strategies do no guarantee any approximation for POMDPs with limsup objective.

4.2. Illustrative Examples

We show an example of a POMDP that will be analyzed in Section 6.2 in light of our technique. This example comes in two variants differing in sophistication.

4.2.1. Simple Version.

Let us explain the easiest version.

Example 4.4. Consider two states (k_u, k_d) and two actions: *up* (a_u) and *down* (a_d). All transitions are possible (including loops) and they do not depend on the action. Signals inform the player when the state changes. In terms of actions and rewards, by playing a_u, the player obtains a reward of 1 only if the current state is k_u. Similarly, by playing a_d, the player obtains a reward of 1 only if the state is k_d.

Figure 5 is a representation of the game with specific transition probabilities.

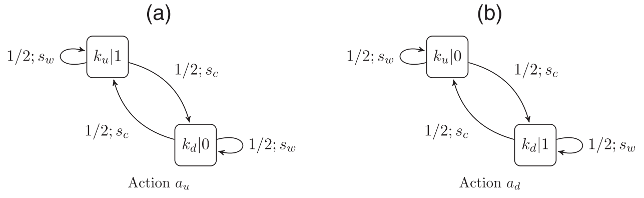
Consider an initial belief $p_1 = 1/4 \cdot \delta_{k_u} + 3/4 \cdot \delta_{k_d}$. During a play, the decision maker can have two beliefs, $1/4 \cdot \delta_{k_u} + 3/4 \cdot \delta_{k_d}$ or $3/4 \cdot \delta_{k_u} + 1/4 \cdot \delta_{k_d}$, because the signals notify when there has been a change. The value of this game is $3/4$. An optimal strategy is to play action a_d until getting a signal s_c, then play action a_u until getting a signal s_c, and repeat.

4.2.2. Involved Version.

Let us go to the more complex version. Now the transition between the two extremes includes more states, instead of being a direct jump.

Example 4.5. Consider six states and four actions: *up* (a_u), *down* (a_d), *left* (a_l), and *right* (a_r). States can be separated into two groups: *extremes* (k_u and k_d) and *transitional* ($k_{l_1}, k_{r_1}, k_{l_2}$, and k_{r_2}). Furthermore, transitional states can be divided into two groups: *left states* (k_{l_1} and k_{l_2}) and *right states* (k_{r_1} and k_{r_2}). Transitions are from extreme states to

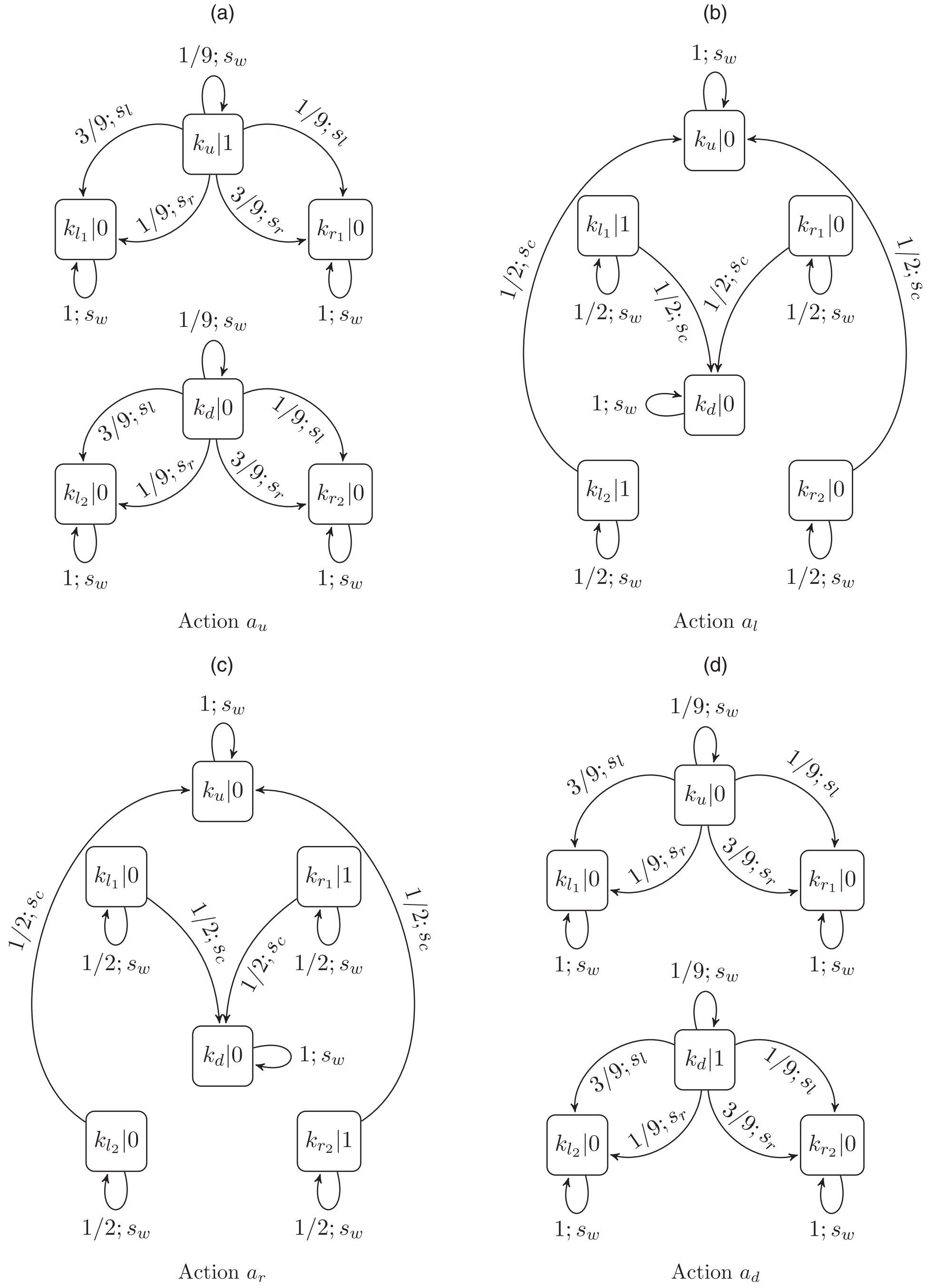
Figure 5. Simple POMDP.



transitional states and from transitional to extremes. More precisely, excluding loops, only the following transitions are possible: from k_u to either k_{l_1} or k_{r_1}, then from these two to k_d, from k_d to either k_{l_2} or k_{r_2}, and then back to k_u. Signals are such that the player knows that (i) the state changed to an extreme state or (ii) the state changed to a transitional state and the new state is with higher probability a left state or a right state. In terms of actions and rewards, each action has an associated set of states in which the reward is 1 and in other states, the reward is 0: by playing a_u, the reward is 1 only if the current state is k_u; playing a_d rewards only state k_d; a_l rewards states k_{l_1} and k_{l_2}; and a_r rewards states k_{r_1} and k_{r_2}. Figure 6 is a representation of this game with specific transition probabilities.

Consider an initial belief $p_1 = 1/4 \cdot \delta_{k_u} + 3/4 \cdot \delta_{k_d}$. The value of the game is 21/32. An optimal strategy is given by playing action a_d until getting a signal s_l or s_r. If the decision maker got signal s_l, then play action a_l;

Figure 6. Complex POMDP.



Action a_u

Action a_l

Action a_r

Action a_d

otherwise, play action a_r. Repeat action a_l or a_r until getting the signal s_c. Then, play a_u until getting a signal s_l or s_r. When this happens, play a_l or a_r accordingly until getting signal s_c. And so repeat the cycle.

The belief dynamic under this optimal strategy is the following. The initial belief is p_1, supported in the extreme states. By getting a signal s_w, the belief does not change. By getting signal s_l, the weight on k_u distributes between states k_{l_1} and k_{r_1} in a $3:1$ proportion, and the weight on k_d distributes between k_{l_2} and k_{r_2} in the same way. By getting signal s_r, the distribution is similar, but the roles of the left states and the right states are interchanged. Once the belief is in the transitional states, by playing the respective action (either a_l or a_r), the belief does not change while receiving signal s_w. Upon receiving the signal s_c, the new belief is $3/4 \cdot \delta_{k_u} + 1/4 \cdot \delta_{k_d}$. By symmetry of the POMDP, the dynamic is then similar until getting signal s_c for a second time. At that time, the belief is equal to the initial distribution, namely, $1/4 \cdot \delta_{k_u} + 3/4 \cdot \delta_{k_d}$.

Remark 4.6. For the decision maker to have a finite-memory strategy, some quantity with finitely many options must be updated over time. A tentative idea is to compute the posterior belief, but it can take infinitely many values. In this example, using a belief partition is enough to encode an optimal strategy. In general, it is an open question whether a belief partition is sufficient to achieve ε-optimal strategies.

5. Structure of the Proof

In this section, we introduce two key lemmas and derive from them the proof of Theorem 2.9. We first define the history at stage m, which is all the information the decision maker has at stage m.

Definition 5.1 (m-Stage History)**.** Given a strategy $\sigma \in \Sigma$ and an initial belief p_1, denote the (random) history at stage m by

$$H_m := ((A_1, S_1), (A_2, S_2), \dots, (A_{m-1}, S_{m-1}))\,.$$

The random variable H_m takes values in $\mathcal{H}_m = (\mathcal{A} \times \mathcal{S})^{m-1}$.

Recall that we denote the state at stage m by K_m, which takes values in $\mathcal{K}$; the signal at stage m by S_m, which takes values in $\mathcal{S}$; and the action at stage m by A_m, which takes values in $\mathcal{A}$. Note that the history at stage m does not contain direct information about the states $K_1, \dots, K_m$.

The belief of the player at stage m plays a key role in the study of POMDPs, and we formally define it as follows.

Definition 5.2 (m-Stage Belief)**.** Given a strategy $\sigma \in \Sigma$ and an initial belief p_1, denote the belief at stage m by P_m, which is given by, for all $k \in \mathcal{K}$,

$$P_m(k) := \mathbb{P}^{p_1}_{\sigma}(K_m = k \mid H_m)\,.$$

For fixed σ and p_1, one can use Bayes' rule to compute P_m. To avoid heavy notations, we omit the dependence of P_m on σ and p_1. For $p \in \Delta(\mathcal{K})$, denote the support of p by $\mathrm{Supp}(p)$, which is the set of $k \in \mathcal{K}$ such that $p(k) > 0$.

In the remainder, the notation "a.s." stands for "almost surely". The first ingredient of the proof of Theorem 2.9 is the following lemma.

Lemma 5.3. *For any initial belief p_1 and $\varepsilon > 0$, there exists $m_\varepsilon \geq 1$, $\sigma^\varepsilon \in \Sigma$, and a (random) belief $P^* \in \Delta(\mathcal{K})$ (which depends on the history before stage m_ε) such that*

1. $$\mathbb{P}^{p_1}_{\sigma^\varepsilon}\left(\|P_{m_\varepsilon} - P^*\|_1 \leq \varepsilon\right) \geq 1 - \varepsilon\,.$$

2. *there exists $\sigma \in \Sigma$, which depends on P^*, such that for all $k \in \mathrm{Supp}(P^*)$,*

$$\left(\frac{1}{n}\sum_{m=1}^{n} G_m\right) \xrightarrow[{[n\to\infty]}]{} \gamma^k_\infty(\sigma) \quad \mathbb{P}^k_\sigma\text{-almost surely}$$

Moreover, $\gamma^{P^*}_\infty(\sigma) = v_\infty(P^*)$ and $\mathbb{E}^{p_1}_{\sigma^\varepsilon}(v_\infty(P^*)) \geq v_\infty(p_1) - \varepsilon$.

This result is a consequence of Venel and Ziliotto [27, lemma 33]. This previous work states the existence of elements $\mu^* \in \Delta(\Delta(\mathcal{K}))$ and $\sigma^* \in \Delta(\Sigma)$ with properties similar to those of $P^* \in \Delta(\mathcal{K})$ and $\sigma \in \Sigma$. In this sense, the present lemma can be seen as a deterministic version of this previous result. To focus on the new tools we introduce in this paper to prove Theorem 2.9, we relegate the proof and the explanation of the differences between the two lemmas to the appendix.

Remark 5.4. The first property of Lemma 5.3 follows immediately from Venel and Ziliotto [27, lemma 33] by the type of convergence in this previous result. On the other hand, the second property requires the introduction of a certain Markov chain on $\mathcal{K} \times \mathcal{A} \times \Delta(\mathcal{K})$. This Markov chain is already present in the work Venel and Ziliotto [27], but was used for other purposes. Therefore, the proof consists mainly of recalling previous results and constructions.

Remark 5.5. Note that $\mathbb{P}_\sigma^k$ represents the law on plays induced by the strategy σ, conditional on the fact that the initial state is k. This does not mean that we consider the decision maker to know k. In the same fashion, $\gamma_\infty^k(\sigma)$ is the reward given by the strategy σ, conditional on the fact that the initial state is k. Even though σ is optimal in $\Gamma(P^*)$, this does not imply that σ is optimal in $\Gamma(\delta_k)$: we may have $\gamma_\infty^k(\sigma) < v_\infty(\delta_k)$.

The importance of Lemma 5.3 comes from the fact that the average rewards converge almost surely (a.s.) to a limit that depends only on the initial state k. Intuitively, this result means that for any initial belief p_1, after a finite number of stages, we can get ε-close to a belief P^* such that the optimal reward from P^* is, in expectation, almost the same as from p_1, and moreover, from P^*, there exists an optimal strategy that induces a strong ergodic behaviour on the state dynamics. Thus, there is a natural way to build a 3ε-optimal strategy $\tilde{\sigma}$ in $\Gamma(p_1)$: first, apply the strategy σ^ε for m_ε stages, then apply σ. Because after m_ε steps the current belief P_{m_ε} is ε-close to P^* with probability higher than $1 - \varepsilon$, the reward from playing $\tilde{\sigma}$ is at least the expectation of $\gamma_\infty^{P^*}(\sigma) - 2\varepsilon$, which is greater than $v_\infty(p_1) - 3\varepsilon$. Therefore, this procedure yields a 3ε-optimal strategy. Nonetheless, σ may not have finite memory, and thus $\tilde{\sigma}$ may not have either. The main difficulty of the proof is to transform σ into a finite-memory strategy. We formalize this discussion below.

Definition 5.6 (Ergodic Strategy)**.** Let $p^* \in \Delta(\mathcal{K})$. We say that a strategy σ is *ergodic* for p^* if the following holds for all $k \in \operatorname{Supp}(p^*)$:

$$\left(\frac{1}{n}\sum_{m=1}^{n} G_m\right) \underset{[n\to\infty]}{\longrightarrow} \gamma_\infty^k(\sigma) \quad \mathbb{P}_\sigma^k\text{-a.s}$$

From the previous discussion, we aim at proving the following result.

Lemma 5.7. *Let $p^* \in \Delta(\mathcal{K})$ and σ be an ergodic strategy for p^*. For all $\varepsilon > 0$, there exists a finite-memory strategy σ' such that*

$$\gamma_\infty^{p^*}(\sigma') \geq \gamma_\infty^{p^*}(\sigma) - \varepsilon\,.$$

This is our key lemma and the main technical contribution. The next section is devoted to explaining the technique used and proving it.

Proof of Theorem 2.9 Assuming Lemmas 5.3 and 5.7. Let p_1 be an initial belief and $\varepsilon > 0$. Let m_ε, σ^ε, P^*, and σ be given by Lemma 5.3. Define the strategy σ^0 by playing σ^ε until stage m_ε, then switch to the strategy σ' given by Lemma 5.7 for σ and $p^* = P^*$. Note that σ^0 has finite memory. We have

$$\begin{aligned}
\gamma_\infty^{p_1}(\sigma^0) &= \mathbb{E}_{\sigma^\varepsilon}^{p_1}\left(\gamma_\infty^{P_{m_\varepsilon}}(\sigma')\right) && \text{(definition of } \sigma^0\text{)}\\
&\geq \mathbb{E}_{\sigma^\varepsilon}^{p_1}\left(\gamma_\infty^{P^*}(\sigma')\right) - 2\varepsilon && \text{(Lemma 5.3)}\\
&\geq \mathbb{E}_{\sigma^\varepsilon}^{p_1}\left(\gamma_\infty^{P^*}(\sigma)\right) - 3\varepsilon && \text{(Lemma 5.7)}\\
&= \mathbb{E}_{\sigma^\varepsilon}^{p_1}(v_\infty(P^*)) - 3\varepsilon && \text{(Lemma 5.3)}\\
&\geq v_\infty(p_1) - 4\varepsilon && \text{(Lemma 5.3)}\,,
\end{aligned}$$

and the theorem is proved. □

6. Supersupport and Proof of Lemma 5.7

In this entire section, fix $p^* \in \Delta(\mathcal{K})$, which will be used as an initial belief, and σ an ergodic strategy for p^*.

6.1. Notation

For $a, b \in \mathbb{R}$, denote the set $[a, b] \cap \mathbb{Z}$ by $[a\,..\,b]$.

Definition 6.1 (Value Partition)**.** Let $\sim$ be the equivalence relationship on $\operatorname{Supp}(p^*)$ defined by $k \sim k'$ if and only if $\gamma_\infty^k(\sigma) = \gamma_\infty^{k'}(\sigma)$. Let $\{\mathcal{K}_1, \ldots, \mathcal{K}_I\}$ be the corresponding *value partition*.

Definition 6.2 (Supersupport)**.** For $i \in [1\,..\,I]$ and $m \geq 0$, define, for all $h_m \in \mathcal{H}_m$,

$$\mathcal{B}^i_m(h_m) := \bigcup_{k_1 \in \mathcal{K}_i} \left\{ k \in \mathcal{K} : \mathbb{P}^{k_1}_\sigma(H_m = h_m) > 0, \mathbb{P}^{k_1}_\sigma(K_m = k | H_m = h_m) > 0 \right\}.$$

In other words, $\mathcal{B}^i_m(h_m)$ is the set of all reachable states at stage m starting from some state in $\mathcal{K}_i$, playing the strategy σ, and obtaining history h_m (if $H_m = h_m$ is possible). Define

$$B^i_m := \mathcal{B}^i_m(H_m),$$

the random set associated with H_m, and $B_m := (B^1_m, \ldots, B^I_m)$, the supersupport at stage m.

Remark 6.3. Note that

$$\operatorname{Supp}(P_m) = \bigcup_{i \in [1\,..\,I]} B^i_m .$$

Therefore, the support of P_m can be deduced from the supersupport B_m. On the other hand, B_m cannot be deduced from P_m, and thus cannot be deduced from the support of P_m. This justifies the vocabulary.

We will build a finite-memory ε-optimal strategy that plays by blocks. Each block has fixed finite length, and, within each block, the strategy depends only on the history in the block and on the supersupport at the beginning of the block. At the end of the block, the automaton computes the new supersupport according to the block history and the previous supersupport. Thus, the only difference with a bounded recall strategy is that our strategy keeps track of the supersupport. Supersupport is a type of origin information: it is related to the value partition, and therefore to where the current mass distribution comes from.

Definition 6.4 (h_m-Shift)**.** Let $m \geq 1$ and $h_m \in \mathcal{H}_m$. The h_m*-shift* of σ is the strategy $\sigma[h_m]$ defined by, for all $m' \geq 1$,

$$\sigma[h_m](h_{m'}) := \sigma(h_m, h_{m'}) .$$

We define $\sigma_m := \sigma[H_m]$, the corresponding random shift at stage m.

In other words, $\sigma[h_m]$ corresponds to the continuation of the strategy σ conditional on the fact that the history of the first m stages was h_m.

6.2. Illustration

The supersupport captures specific information related to the beginning of the game: the origin of the current mass distribution (given by P_m) in terms of the initial value partition $(\mathcal{K}_i)_{i \in [1\,..\,I]}$. There are finitely many possible supersupports, and it is possible to keep track of the current supersupport using Bayesian updating. Therefore, it is a good variable to be used in finite-memory strategies.

Let us recall our simple example of a POMDP, Example 4.4.

Example 4.4 (Restated)**.** Consider two states (k_u, k_d) and two actions: *up* (a_u) and *down* (a_d). All transitions are possible (including loops), and they do not depend on the action. Signals inform the player when the state changes. In terms of actions and rewards, by playing a_u, the player obtains a reward of 1 only if the current state is k_u. Similarly, by playing a_d, the player obtains a reward of 1 only if the state is k_d.

Finite recall is enough to approximate the value of this POMDP: the decision maker can recall the last action. Then, upon seeing the signal s_c, the player has to change actions. Recall that $p_1 = 1/4 \cdot \delta_{k_u} + 3/4 \cdot \delta_{k_d}$. Therefore, an optimal strategy is given by playing a_d until getting signal s_c, then playing a_u until getting signal s_c and repeating. This strategy is ergodic for p_1, and the corresponding value partition is given by $(\mathcal{K}_1 = \{k_u\}$, $\mathcal{K}_2 = \{k_d\})$, because if $K_1 = k_u$, the long-run reward is 0, and if $K_1 = k_d$, the long-run reward is 1. In this case, the supersupport describes completely the belief P_m because it keeps track of which state has the highest (or lowest) probability.

Although the example is simple, we can already see the difference between support strategies and supersupport strategies. In this case, all strategies based on the current support (the support of P_m) are constant and therefore can achieve a long-run reward of at most 1/2. On the other hand, supersupport strategies can be optimal and achieve a long-run reward of 3/4.

This example also shows that playing by blocks and defining the behaviour in each block by the current support (instead of the supersupport) is not enough.

Let us analyze now our more complex POMDP example, Example 4.5.

Example 4.5 (Restated)**.** Consider six states and four actions: *up* (a_u), *down* (a_d), *left* (a_l), and *right* (a_r). States can be separated into two groups: *extremes* (k_u and k_d) and *transitional* ($k_{l_1}, k_{r_1}, k_{l_2}$, and k_{r_2}). Furthermore, transitional states can be divided into two groups: *left states* (k_{l_1} and k_{l_2}) and *right states* (k_{r_1} and k_{r_2}). Transitions are from extreme states to transitional states and from transitional to extremes. More precisely, excluding loops, only the following transitions are possible: from k_u to either k_{l_1} or k_{r_1}, then from these two to k_d, from k_d to either k_{l_2} or k_{r_2}, and then back to k_u. Signals are such that the player knows that (i) the state changed to an extreme state or (ii) the state changed to a transitional state and the new state is with higher probability a left state or a right state. In terms of actions and rewards, each action has an associated set of states in which the reward is 1 and the rest are 0: by playing a_u, the reward is 1 only if the current state is k_u; playing a_d rewards only state k_d; a_l rewards states k_{l_1} and k_{l_2}; and a_r rewards states k_{r_1} and k_{r_2}. Figure 6 is a representation of this game with specific transition probabilities.

Recall that $p_1 = 1/4 \cdot \delta_{k_u} + 3/4 \cdot \delta_{k_d}$ and that an optimal strategy is given by playing action a_d until getting a signal s_l or s_r. If the decision maker gets signal s_l, then play action a_l, otherwise, play action a_r. Repeat action a_r until getting the signal s_c. Then, play a_u until getting a signal s_l or s_r. When this happens, play a_l or a_r accordingly until getting signal s_c. And so repeat the cycle.

This optimal strategy is ergodic for p_1, and the corresponding value partition is given by ($\mathcal{K}_1 = \{k_u\}, \mathcal{K}_2 = \{k_d\}$), because if $K_1 = k_u$, the long-run reward is 0, and if $K_1 = k_d$, the long-run reward is 7/8. Contrary to the previous example, the supersupport does not describe completely the belief P_m. Indeed, consider the initial belief p_1, which is supported on the extreme states, and that the decision maker gets either signal s_l or s_r. Then, the new belief is supported in all the transitional states, and the supersupport is the same under any of these two histories and equal to $B = (B^1 = \{k_{l_1}, k_{r_1}\}, B^2 = \{k_{l_2}, k_{r_2}\})$. Based on this supersupport one cannot reconstruct the current belief, but one knows more than only the support: we can differentiate the origin (k_u or k_d) of the current belief distribution.

Notice that using the supersupport alone is not enough to get ε-optimal strategies. Indeed, in transitional states, the decision maker needs to know whether the state is more likely to be in a left state or a right state in order to play well, and the supersupport does not contain such information. That is why, in the proof of Lemma 5.7, we consider a more sophisticated class of strategies that combine supersupport and bounded recall. For the moment, let us describe such a strategy for this example. Choose n_0 very large, and for each $\ell \geq 1$, play the following strategy in the time block $[\ell n_0 + 1 \,..\, (\ell+1)n_0]$:

Case 1. *The supersupport at stage* $\ell n_0 + 1$ *is* $(\{k_u\}, \{k_d\})$. Play the previous 0-optimal strategy; that is, play action a_d until getting a signal s_l or s_r. If the decision maker gets signal s_l, then play action a_l; otherwise, play action a_r. Repeat action a_r until getting the signal s_c. Then, play a_u until getting a signal s_l or s_r. When this happens, play a_l or a_r accordingly until getting signal s_c. And so repeat the cycle.

Case 2. *The supersupport at stage* $\ell n_0 + 1$ *is* $(\{k_d\}, \{k_u\})$. Play the same strategy as in Case 1, except that the roles of a_u and a_d are switched.

Case 3. *The supersupport at stage* $\ell n_0 + 1$ *is* $(\{k_{l_1}, k_{r_1}\}, \{k_{l_2}, k_{r_2}\})$. Play a_l (or a_r) until getting the signal s_c. At this point, the supersupport is $(\{k_d\}, \{k_u\})$. Then, play as in Case 2.

Case 4. *The supersupport at stage* $\ell n_0 + 1$ *is* $(\{k_{l_2}, k_{r_2}\}, \{k_{l_1}, k_{r_1}\})$. Play a_l (or a_r) until getting the signal s_c. At this point, the supersupport is $(\{k_u\}, \{k_d\})$. Then, play as in Case 1.

This strategy is suboptimal during the first phase of Case 3 and Case 4, until the decision maker receives signal s_c. As n_0 grows larger and larger, this part becomes negligible. Thus, for any $\varepsilon > 0$, there exists n_0 such that this strategy is ε-optimal (but not optimal).

6.3. Properties

Now we can state properties of supersupports when the strategy σ is ergodic for p^* and explain how rich is the structure of the random sequence of beliefs $(P_m)_{m \geq 1}$. By definition of ergodic strategies, the map $k \mapsto \gamma_\infty^k(\sigma)$ is constant on $\mathcal{K}_i$, and we denote its value by γ_∞^i.

Lemma 6.5 (Continuation Value)**.** *For all* $m \geq 1$ *and* $i \in [1 \,..\, I]$, *it holds* $\mathbb{P}_\sigma^{p^*}$*-a.s. that*

$$\forall k \in B_m^i \quad \gamma_\infty^k(\sigma_m) = \gamma_\infty^i$$

Consequently, $B_m^1, \ldots, B_m^I$ *are disjoint* $\mathbb{P}_\sigma^{p^*}$*-a.s.*

Proof. Let $i \in [1 \,..\, I]$. Considering the law given by $\mathbb{P}_\sigma^{p^*}$, fix a realization $K_m = k \in B_m^i$. By definition of supersupport, there exists $k' \in \mathcal{K}_i \subseteq \operatorname{Supp}(p^*)$ such that k can be reached from k' in m steps. Recall that because σ is ergodic for p^*,

$$\left(\frac{1}{n} \sum_{m'=m}^{m+n-1} G_{m'} \right) \xrightarrow[{[n \to \infty]}]{} \gamma_\infty^{k'}(\sigma) = \gamma_\infty^i \quad \mathbb{P}_\sigma^{k'}\text{-a.s.}$$

In particular, the convergence holds when $K_m = k$. Then,

$$\left(\frac{1}{n}\sum_{m'=1}^{n} G_{m'}\right) \xrightarrow[{[n\to\infty]}]{} \gamma^{k'}_{\infty}(\sigma) = \gamma^{i}_{\infty} \quad \mathbb{P}^{k}_{\sigma_m}\text{-}a.s.\,,$$

and therefore $\gamma^{k}_{\infty}(\sigma_m) = \gamma^{i}_{\infty}$. □

Another property of the supersupport is concerned with consecutive conditioning and is fairly intuitive. We formally state it in the following lemma and show the proof for completeness.

Lemma 6.6 (Continuation Supersupport)**.** *Let $i \in [1\,..\,I]$, $m, m' \geq 0$. For all realizations $H_{m+m'} = h = (h_m, h_{m'}) \in \mathcal{H}_{m+m'}$, denoting $\mathcal{C} = \mathcal{B}^{i}_{m+m'}(h)$, we have that, for all $k \in \mathcal{B}^{i}_{m}(h_m)$,*

$$\mathbb{P}^{k}_{\sigma[h_m]}(K_{m'} \in \mathcal{C} \,|\, H_{m'} = h_{m'}) = 1\,.$$

In other words, the supersupport that arises at stage $m + m'$, $B_{m+m'}$, coincides with the supersupport that would arise from a two-step procedure: first advancing m stages, and then applying the continuation of the strategy, σ_m, for m' more stages.

Proof. Fix a realization $H_{m+m'} = h = (h_m, h_{m'})$, and let $k \in \mathcal{B}^{i}_{m}(h_m)$. Recall that, by definition of supersupport,

$$\mathcal{B}^{i}_{m}(h_m) = \bigcup_{\substack{\bar{k}_1 \in \mathcal{K}_i \\ \mathbb{P}^{\bar{k}_1}_{\sigma}(h_m > 0)}} \operatorname{Supp}\left(\mathbb{P}^{\bar{k}_1}_{\sigma}(K_m = \cdot \,|\, H_m = h_m)\right).$$

Therefore, there exists $\bar{k}_1 \in \mathcal{K}_i$ such that $k \in \operatorname{Supp}(\mathbb{P}^{\bar{k}_1}_{\sigma}(K_m = \cdot \,|\, H_m = h_m))$. In particular, we have that $\mathbb{P}^{\bar{k}_1}_{\sigma}(K_m = k) > 0$.

Consider k' such that $\mathbb{P}^{k}_{\sigma[h_m]}(K_{m'} = k' \,|\, H_{m'} = h_{m'}) > 0$. By a semigroup property, we deduce that

$$\mathbb{P}^{\bar{k}_1}_{\sigma}(K_{m+m'} = k' \,|\, H_{m+m'} = h) > 0\,,$$

which implies that $k' \in \mathcal{B}^{i}_{m+m'}(h) = \mathcal{C}$, and thus the lemma is proved. □

Remark 6.7. This property does not depend on the fact that σ is ergodic for p^*.

6.4. Proof of Lemma 5.7

Fix $p^* \in \Delta(\mathcal{K})$ such that σ is ergodic for p^*. Note that, for all $m \geq 1$, $B_m \in \{(\mathcal{C}_1, \ldots, \mathcal{C}_I) : \mathcal{C}_1, \ldots, \mathcal{C}_I \subseteq \mathcal{K}\}$, which is a finite set. Denote all different supersupports that can occur with positive probability by $\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^J$, that is,

$$\left\{\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^J\right\} := \bigcup_{m\geq 1} \operatorname{Supp} B_m.$$

Moreover, because $\mathcal{D}^j$ corresponds to a supersupport that occurs at some stage and under some history, there exists h^j and m_j such that $h^j \in \mathcal{H}_{m_j}$ and $\mathcal{B}^{i}_{m_j}(h^j)$. In other words, $\mathcal{D}^j$ is the realization of the supersupport at stage m_j under history h^j, and $\{\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^J\}$ contains all supersupports that can occur.

6.4.1. Definition of the strategy σ'. Let $\varepsilon > 0$. By Lemma 6.5, there exists $n_0 \in \mathbb{N}^*$ such that for all $i \in [1\,..\,I]$, $j \in [1\,..\,J]$, and $k \in \mathcal{D}^j_i$,

$$\mathbb{E}^{k}_{\sigma[h^j]}\left(\frac{1}{n_0}\sum_{m=1}^{n_0} G_m\right) \geq \gamma^{i}_{\infty} - \varepsilon.$$

Define the strategy σ' by blocks, and characterize each block by induction. For each $\ell \geq 0$, the block number ℓ consists in the stages m such that $\ell n_0 + 1 \leq m \leq (\ell + 1)n_0$. We characterize the behaviour in block ℓ by a variable $J_\ell \in [1\,..\,J]$ in the following way. For stage m inside block ℓ, the strategy σ' plays according to J_ℓ and the history between stages $\ell n_0 + 1$ and m. Each block is characterized by induction because the variable J_ℓ is computed at stage $\ell n_0 + 1$ according to $J_{(\ell-1)}$ and the history in the last n_0 stages. Thus, σ' can be seen as mapping from $\cup_{m=1}^{n_0} \mathcal{H}_m \times [1\,..\,J]$ to $\mathcal{A}$.

Consider $\ell = 0$, that is, the first block. The strategy σ' is defined on the first n_0 stages as follows. Consider the value partition $\{\mathcal{K}_1, \dots, \mathcal{K}_I\}$ given by p^* and σ. By definition of $\mathcal{D}^1, \mathcal{D}^2, \dots, \mathcal{D}^J$, there exists $j \in [1 \,..\, J]$ such that $B_1 = \mathcal{D}^j$. Set $J_0 = j$, and define $\sigma'(h, J_0) := \sigma(h^{J_0}, h)$ for all $h \in \mathcal{H}_m$ and $m \le n_0$.

Let us proceed to the induction step. Consider $\ell \ge 1$ and assume that we have defined $J_{(\ell-1)}$ and σ' up to stage ℓn_0. Denote the history between stages $(\ell - 1)n_0 + 1$ and $\ell n_0 + 1$ by $h \in \mathcal{H}_{n_0+1}$, and define J_ℓ such that, for all $i \in [1 \,..\, I]$,

$$\mathcal{D}_i^{J_\ell} = \mathcal{B}^i_{m_{J_{(\ell-1)}} + n_0 + 1}\left(h^{J_{(\ell-1)}}, h\right).$$

Then, extend σ' for n_0 additional stages as before: $\sigma'(h, J_\ell) := \sigma(h^{J_\ell}, h)$ for all $h \in \mathcal{H}_m$ and $m \le n_0$. Thus, we have defined J_ℓ and extended σ' up to stage $(\ell + 1)n_0$.

To summarize our construction in words, during stages $\ell n_0 + 1, \ell n_0 + 2, \dots, (\ell + 1)n_0$, the decision maker plays as if he was playing σ from history h^{J_ℓ}. Notice that the indexes $J_0, J_1, \dots$ depend on the history, and therefore are random. Now we will connect the strategy σ' with the supersupport given by p^* and σ.

Lemma 6.8. *For all $i \in [1 \,..\, I]$, $k \in \mathcal{K}_i$, and $\ell \ge 0$, we have that*

$$\mathbb{P}^k_{\sigma'}\left(K_{\ell n_0 + 1} \in \mathcal{D}_i^{J_\ell}\right) = 1.$$

Consequently, $\mathcal{D}^{J_\ell} = (\mathcal{D}_1^{J_\ell}, \mathcal{D}_2^{J_\ell}, \dots, \mathcal{D}_I^{J_\ell})$ is a partition of the support of $P_{\ell n_0 + 1}$. Moreover,

$$\mathbb{P}^{p^*}_{\sigma'}\left(K_{\ell n_0 + 1} \in \mathcal{D}_i^{J_\ell}\right) = p^*(\mathcal{K}_i).$$

Proof. Fix $i \in [1 \,..\, I]$ and $k \in \mathcal{K}_i$. We will prove the result by induction on $\ell \ge 0$. For $\ell = 0$, $\mathcal{D}^{J_0} = (\mathcal{K}_1, \dots, \mathcal{K}_I)$, and $\mathbb{P}^k_{\sigma'}(K_1 \in \mathcal{D}_i^{J_0} = \mathcal{K}_i) = 1$. Thus, the result holds.

Assume $\ell \ge 1$. Note that $H_{(\ell-1)n_0+1}$ determines the value of $J_0, \dots, J_{(\ell-1)}$. Therefore, $\mathcal{D}_i^{J_{(\ell-1)}}$ is also determined by $H_{(\ell-1)n_0+1}$. By induction hypothesis,

$$\mathbb{P}^k_{\sigma'}\left(K_{(\ell-1)n_0+1} \in \mathcal{D}_i^{J_{(\ell-1)}}\right) = 1.$$

We must prove that, under these circumstances, $K_{\ell n_0 + 1} \in \mathcal{D}_i^{J_\ell}$.

Indeed, index $J_{\ell-1}$ defines strategy σ' for stages $\ell n_0 + 1, \ell n_0 + 2, \dots, (\ell + 1)n_0$: the decision maker will play according to $\sigma[h^{J_{(\ell-1)}}]$. By playing σ' during this block, a history $h \in \mathcal{H}_{n_0+1}$ will be collected. Let $m := J_{(\ell-1)}$ and $m' := n_0 + 1$. By Lemma 6.8, we have that starting from $K_{(\ell-1)n_0+1} \in \mathcal{D}_i^{J_{(\ell-1)}}$, playing $\sigma[h^{J_{(\ell-1)}}]$ during n_0 stages, and collecting history $h \in \mathcal{H}_{n_0+1}$ leads to a state that, by definition of J_ℓ, is in $\mathcal{D}_i^{J_\ell}$. Therefore,

$$\mathbb{P}^k_{\sigma'}\left(K_{\ell n_0 + 1} \in \mathcal{D}_i^{J_\ell}\right) \ge \mathbb{P}^k_{\sigma'}\left(K_{(\ell-1)n_0+1} \in \mathcal{D}_i^{J_{(\ell-1)}}\right) = 1,$$

which proves the first result.

Now we know that the union of $\mathcal{D}_1^{J_\ell}, \mathcal{D}_2^{J_\ell}, \dots, \mathcal{D}_I^{J_\ell}$ covers the support of $\mathbb{P}^{p^*}_{\sigma'}(K_{\ell n_0 + 1} = \cdot)$. Moreover, by Lemma 6.5, $\mathcal{D}_1^{J_\ell}, \mathcal{D}_2^{J_\ell}, \dots, \mathcal{D}_I^{J_\ell}$ are disjoint. Because $(\mathcal{K}_1, \dots, \mathcal{K}_I)$ partitions the support of p^*, we have that, for all $i \in [1 \,..\, I]$,

$$\begin{aligned}
\mathbb{P}^{p^*}_{\sigma'}\left(K_{\ell n_0 + 1} \in \mathcal{D}_i^{J_\ell}\right) &= \sum_{i'=1}^{I} \sum_{k \in \mathcal{K}_{i'}} p^*(k) \mathbb{P}^k_{\sigma'}\left(K_{\ell n_0 + 1} \in \mathcal{D}_i^{J_\ell}\right) \\
&= \sum_{i'=1}^{I} \sum_{k \in \mathcal{K}_{i'}} p^*(k) \mathbb{1}_{k \in \mathcal{K}_i} \\
&= p^*(\mathcal{K}_i),
\end{aligned}$$

which proves the second property. □

To finish the proof of Lemma 5.7, we must show that the finite-memory strategy σ' guarantees the reward obtained by σ up to ε. The idea is that in each block we are playing some shift of σ for n_0 stages. The shift is chosen so that information about the initial belief is correctly updated, whereas the number n_0 is chosen so that the expected average reward of the whole block is close to the expected limit average reward. Then, because all blocks have the same approximation error, the average considering all blocks yields approximately $\gamma^{p^*}_\infty(\sigma)$. This is the intuition behind the following lemma.

Lemma 6.9. *Let $L \in \mathbb{N}^*$. The following inequality holds:*

$$\mathbb{E}_{\sigma'}^{p^*}\left(\frac{1}{Ln_0}\sum_{m=1}^{Ln_0} G_m\right) \geq \gamma_\infty^{p^*}(\sigma) - \varepsilon.$$

Proof. We have, for all $\ell \geq 0$,

$$\begin{aligned}
\mathbb{E}_{\sigma'}^{p^*}\left(\tfrac{1}{n_0}\sum_{m=\ell n_0+1}^{(\ell+1)n_0} G_m\right) &= \mathbb{E}_{\sigma'}^{p^*}\left(\mathbb{E}_{\sigma[h^{J_\ell}]}^{K_{\ell n_0+1}}\left(\tfrac{1}{n_0}\sum_{m=1}^{n_0} G_m\right)\right) && \text{(definition of } \sigma'\text{)}\\
&= \mathbb{E}_{\sigma'}^{p^*}\left(\sum_{i=1}^{I}\sum_{k\in\mathcal{D}_i^{J_\ell}} \mathbb{P}_{\sigma'}^{p^*}(K_{\ell n_0+1} = k)\mathbb{E}_{\sigma[h^{J_\ell}]}^{k}\left(\frac{1}{n_0}\sum_{m=1}^{n_0} G_m\right)\right) && \text{(Lemma 6.8)}\\
&\geq \mathbb{E}_{\sigma'}^{p^*}\left(\sum_{i=1}^{I}\sum_{k\in\mathcal{D}_i^{J_\ell}} \mathbb{P}_{\sigma'}^{p^*}(K_{\ell n_0+1} = k)\left[\gamma_\infty^{i}(\sigma) - \varepsilon\right]\right) && \text{(definition of } n_0\text{)}\\
&= \left(\sum_{i=1}^{I} \mathbb{P}_{\sigma'}^{p^*}\left(K_{\ell n_0+1} \in \mathcal{D}_i^{J_\ell}\right)\left[\gamma_\infty^{i}(\sigma) - \varepsilon\right]\right)\\
&= \sum_{i=1}^{I} p^*(\mathcal{K}_i)\left[\gamma_\infty^{i}(\sigma) - \varepsilon\right] && \text{(Lemma 6.8)}\\
&= \gamma_\infty^{p^*}(\sigma) - \varepsilon\,.
\end{aligned}$$

It follows that

$$\mathbb{E}_{\sigma'}^{p^*}\left(\frac{1}{Ln_0}\sum_{m=1}^{Ln_0} G_m\right) = \frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}_{\sigma'}^{p^*}\left(\frac{1}{n_0}\sum_{m=\ell n_0+1}^{(\ell+1)n_0} G_m\right) \geq \gamma_\infty^{p^*}(\sigma) - \varepsilon\,. \quad \square$$

To conclude, because σ' has finite memory, we have

$$\lim_{L\to+\infty}\mathbb{E}_{\sigma'}^{p^*}\left(\frac{1}{Ln_0}\sum_{m=1}^{Ln_0} G_m\right) = \mathbb{E}_{\sigma'}^{p^*}\left(\liminf_{n\to+\infty}\frac{1}{n}\sum_{m=1}^{n} G_m\right) = \gamma_\infty^{p^*}(\sigma')\,,$$

and the above lemma implies that $\gamma_\infty^{p^*}(\sigma') \geq \gamma_\infty^{p^*}(\sigma) - \varepsilon$, which proves Lemma 5.7: for each ergodic strategy σ and $\varepsilon > 0$, one can construct a finite-memory strategy σ' that guarantees the reward obtained by σ up to ε.

Acknowledgments

The authors are grateful to the associate editor and one anonymous referee for their careful rereading and insightful comments. They thank Xavier Venel for fruitful discussions.

Appendix. Proof of Lemma 5.3

A.1. Notation

Recall that Lemma 5.3 is a consequence of Venel and Ziliotto [27, lemma 33]. Thus, we start by introducing some of the terms used in Venel and Ziliotto [27], namely, the n-stage game, invariant measure, occupation measure, and Kantorovich–Rubinstein distance.

Definition A.1 (n-Stage Game). Given a POMDP $\Gamma = (\mathcal{K}, \mathcal{A}, \mathcal{S}, q, g)$, we denote by Γ_n the n-stage game with a value defined by

$$v_n(p) := \sup_{\sigma\in\Sigma}\gamma_n^p(\sigma)\,,$$

where $\gamma_n^p(\sigma) := n^{-1}\mathbb{E}_\sigma^p(\sum_{m=1}^n G_m)$.

Remark A.2. As the notation suggests, it was proven in Venel and Ziliotto [27] that for any finite POMDP, $(v_n)\xrightarrow[n\to\infty]{} v_\infty$ uniformly. The fact that $(v_n)_{n\geq 1}$ converges was proven in Rosenberg et al. [23].

The set $\Delta(\mathcal{K})$ is equipped with its Borelian σ-algebra $\mathcal{B}(\Delta(\mathcal{K}))$, and $\mathcal{C}(\Delta(\mathcal{K}), [0,1])$ denotes the set of continuous functions from $\Delta(\mathcal{K})$ to $[0,1]$.

Definition A.3 (Invariant Measure)**.** Given a POMDP $\Gamma = (\mathcal{K}, \mathcal{A}, \mathcal{S}, q, g)$, $\mu \in \Delta(\Delta(\mathcal{K}))$, and $\sigma : \Delta(\mathcal{K}) \to \Delta(\mathcal{A})$ measurable, we say that μ is σ-invariant if, $\forall f \in \mathcal{C}(\Delta(\mathcal{K}), [0,1])$, we have that

$$\int_{\Delta(\mathcal{K})} \mathbb{E}[f(\tilde{q}[p, \sigma(p)])]\mu(dp) = \int_{\Delta(\mathcal{K})} f(p)\mu(dp)\,,$$

where $\tilde{q} : \Delta(\mathcal{K}) \times \mathcal{A} \to \Delta(\Delta(\mathcal{K}))$ is the natural transition in $\Delta(\mathcal{K})$ from one belief to another, given by Bayes' rule.

The above definition can be intuitively understood in the following way: if the initial belief is distributed according to μ, and the decision maker plays the stationary strategy σ at stage 1, then the belief at stage 2 is distributed according to μ too.

Remark A.4. Because $v_\infty : \Delta(\mathcal{K}) \to [0,1]$ is a continuous function, one can replace f by v_∞ in the previous definition. Moreover, interpreting σ as a (mixed) stationary strategy, we would have that the sequence $(\mathbb{E}_\sigma^\mu[v_\infty(P_m)])_{m\geq 1}$ is constant.

Definition A.5 (*m*-Stage Occupation Measure)**.** Given a POMDP $\Gamma = (\mathcal{K}, \mathcal{A}, \mathcal{S}, q, g)$, a measure $\mu \in \Delta(\Delta(\mathcal{K}))$, and a strategy σ, consider the following dynamic over $\Delta(\mathcal{K})$. First, P_1 is drawn according to μ. Then, $(P_n)_{n\geq 1}$ is obtained by playing according to σ. This way, for each $m \geq 1$, we have that Γ, μ, and σ induce a probability over $\Delta(\mathcal{K})$: for each measurable set $\mathcal{A} \subseteq \Delta(\mathcal{K})$, we can define $\mathbb{P}_\sigma^\mu(P_m \in \mathcal{A})$. Therefore, the m-stage belief, P_m, is a random belief.

We denote the m-stage occupation measure $z_m[\mu, \sigma] \in \Delta(\Delta(\mathcal{K}))$ by the law of P_m over $\Delta(\mathcal{K})$. Formally, $z_m[\mu, \sigma] : \mathcal{B}(\Delta(\mathcal{K})) \to [0,1]$ is given by, for all $\mathcal{C} \in \mathcal{B}(\Delta(\mathcal{K}))$,

$$z_m[\mu, \sigma](\mathcal{C}) = \mathbb{P}_\sigma^\mu(P_m \in \mathcal{C})\,.$$

For sake of notation, we identify $\Delta(\mathcal{K})$ with the extreme points of $\Delta(\Delta(\mathcal{K}))$.

Definition A.6 (Kantorovich–Rubinstein Distance)**.** For all $z, z' \in \Delta(\Delta(\mathcal{K}))$, define

$$d_{KR}(z, z') := \sup_{f \in \mathcal{E}_1} \left| \int_{\Delta(\mathcal{K})} f(p)z(dp) - \int_{\Delta(\mathcal{K})} f(p)z'(dp) \right|,$$

where $\mathcal{E}_1$ is the set of 1-Lipschitz functions from $\Delta(\mathcal{K})$ to $[0,1]$.

Remark A.7. The set $\Delta(\Delta(\mathcal{K}))$ equipped with distance d_{KR} is a compact metric space.

A.2. Proof

Now we can state Venel and Ziliotto [27, lemma 33].

Lemma A.8. *Consider a POMDP Γ, and let $p_1 \in \Delta(\mathcal{K})$. There exists $\mu^* \in \Delta(\Delta(\mathcal{K}))$ and a (mixed) stationary strategy $\sigma^* : \Delta(\mathcal{K}) \to \Delta(\mathcal{A})$ such that*

1. *μ^* is σ^*-invariant;*
2. *for all $\varepsilon > 0$ and $N \geq 1$, there exists $n_\varepsilon \geq N$ and σ^ε a pure strategy in Γ such that σ^ε is 0-optimal in the n_ε-stage game $\Gamma_{n_\varepsilon}(p_1)$ and*

$$d_{KR}\left(\frac{1}{n_\varepsilon}\sum_{m=1}^{n_\varepsilon} z_m[p_1, \sigma^\varepsilon], \mu^*\right) \leq \varepsilon.$$

3.

$$\int_{\Delta(\mathcal{K})} g(p, \sigma^*(p))\mu^*(dp) = \int_{\Delta(\mathcal{K})} v_\infty(p)\mu^*(dp) = v_\infty(p_1).$$

Remark A.9. Lemma A.8 works with elements in $\Delta(\Delta(\mathcal{K}))$ and a mixed stationary strategy, whereas Lemma 5.3 deals with (random) elements in $\Delta(\mathcal{K})$ and a pure strategy. In this sense, we would like to "go down a level": move from μ^* to a random P^*, from z_m to P_m, and still preserve a relationship between $v_\infty(p_1)$ and $\mathbb{E}_{\sigma^\varepsilon}^{p_1}(v_\infty(P^*))$. The ergodic property 2 of Lemma 5.3 follows from the first and third properties of Lemma A.8.

Proof of Lemma 5.3. Consider $p_1 \in \Delta(\mathcal{K})$ and $\varepsilon > 0$ fixed. Because $(v_n)_{n\geq 1}$ converges uniformly to v_∞, consider $N \geq 1$ such that $\varepsilon \geq 1/N$ and, $\forall n \geq N$, $\|v_n - v_\infty\|_\infty \leq \varepsilon$. Now, using Lemma A.8, there exists μ^* and σ^* such that μ^* is σ^*-invariant and, considering ε^4, $\exists n_\varepsilon \geq N$ such that

$$d_{KR}\left(\frac{1}{n_\varepsilon}\sum_{m=1}^{n_\varepsilon} z_m[p_1, \sigma^\varepsilon], \mu^*\right) \leq \varepsilon^4,$$

with $\sigma^\varepsilon \in \Gamma_{n_\varepsilon}(p_1)$ an optimal pure strategy for the n_ε-stage game starting in p_1.

We claim that $\exists m_\varepsilon \le \lceil \varepsilon n_\varepsilon \rceil$ such that

$$\mathbb{P}^{p_1}_{\sigma^\varepsilon}\left(P_{m_\varepsilon} \in \operatorname{Supp}(\mu^*) + B(0,\varepsilon)\right) > 1 - \varepsilon. \tag{A.1}$$

Proceeding by contradiction, assume that $\forall m \le \lceil \varepsilon n_\varepsilon \rceil$, we have $\mathbb{P}^{p_1}_{\sigma^\varepsilon}(P_m \in \operatorname{Supp}(\mu^*) + B(0,\varepsilon)) \le 1 - \varepsilon$. Define the function $f : \Delta(\Delta(\mathcal{K})) \to [0,1]$ by $f(p) = d_\infty(p, \operatorname{Supp}(\mu^*))$, the supremum distance from $\operatorname{Supp}(\mu^*)$. Clearly, $f \in \mathcal{E}_1$. Moreover,

$$\begin{aligned}
\varepsilon^4 &\ge d_{KR}\left(\tfrac{1}{n_\varepsilon}\sum_{m=1}^{n_\varepsilon} z_m[p_1,\sigma^\varepsilon], \mu^*\right) \\
&\ge \left| \int_{\Delta(\mathcal{K})} f(p) \frac{1}{n_\varepsilon}\sum_{m=1}^{n_\varepsilon} z_m[p_1,\sigma^\varepsilon](dp) - \int_{\Delta(\mathcal{K})} f(p)\mu^*(dp) \right| \\
&\ge \left| \tfrac{1}{n_\varepsilon}\sum_{m=1}^{n_\varepsilon} \int_{\Delta(\mathcal{K})} f(p) z_m[p_1,\sigma^\varepsilon](dp) \right| && \left(f(p) = 0, p \in \operatorname{Supp}(\mu^*)\right) \\
&\ge \tfrac{1}{n_\varepsilon}\sum_{m=1}^{\lceil \varepsilon n_\varepsilon \rceil} \int_{\Delta(\mathcal{K})\setminus\left(\operatorname{Supp}(\mu^*)+B(0,\varepsilon)\right)} f(p) z_m[p_1,\sigma^\varepsilon](dp) && \left(f, z_m[p_1,\sigma^\varepsilon] \ge 0\right) \\
&\ge \varepsilon \tfrac{1}{n_\varepsilon}\sum_{m=1}^{\lceil \varepsilon n_\varepsilon \rceil} z_m[p_1,\sigma^\varepsilon](\Delta(\mathcal{K}) \setminus \left(\operatorname{Supp}(\mu^*) + B(0,\varepsilon)\right) && (\text{definition of } f) \\
&\ge \varepsilon^3 && (\text{contradiction hypothesis}),
\end{aligned}$$

which is a contradiction for $\varepsilon < 1$. Thus, we have proven (A.1).

Take $P^* \in \operatorname{argmin}_{p \in \operatorname{Supp}(\mu^*)} \|p - P_{m_\varepsilon}\|_1$. By Equation (A.1), the first property of Lemma 5.3 is satisfied.

For the second property of Lemma 5.3, note that, with probability higher than $1 - \varepsilon$,

$$\begin{aligned}
v_\infty(P^*) &\ge v_{n_\varepsilon}(P^*) - \varepsilon && (\|v_{n_\varepsilon} - v_\infty\|_\infty \le \varepsilon) \\
&\ge v_{n_\varepsilon}\left(P_{m_\varepsilon}\right) - 2\varepsilon && (v_{n_\varepsilon} \text{ is 1-Lipschitz}).
\end{aligned}$$

On the other hand, taking expectation, we get that

$$\begin{aligned}
\mathbb{E}^{p_1}_{\sigma^\varepsilon}\left(v_{n_\varepsilon}\left(P_{m_\varepsilon}\right)\right) &\ge \mathbb{E}^{p_1}_{\sigma^\varepsilon}\left(v_{n_\varepsilon - m_\varepsilon}\left(P_{m_\varepsilon}\right)\right) - \varepsilon && (m_\varepsilon \le \lceil \varepsilon n_\varepsilon \rceil) \\
&\ge v_{n_\varepsilon}(p_1) - 2\varepsilon && \left(\sigma^\varepsilon \text{ is 0-optimal in } \Gamma_{n_\varepsilon}(p_1)\right) \\
&\ge v_\infty(p_1) - 3\varepsilon && (\forall n \ge N, \quad \|v_n - v_\infty\|_\infty \le \varepsilon).
\end{aligned}$$

Therefore, we conclude that

$$\mathbb{E}^{p_1}_{\sigma^\varepsilon}\left(v_\infty\left(P^*\right)\right) \ge v_\infty(p_1) - 6\varepsilon.$$

To complete the proof of Lemma 5.3, given P^*, we need a (pure) strategy σ such that

$$\forall\ k \in \operatorname{Supp}(P^*) \quad \left(\frac{1}{n}\sum_{m=1}^{n} G_m\right) \xrightarrow[{[n\to\infty]}]{} \gamma^k_\infty(\sigma) \quad \mathbb{P}^k_\sigma\text{-a.s.} \tag{A.2}$$

and such that

$$\gamma^{P^*}_\infty(\sigma) = v_\infty(P^*). \tag{A.3}$$

Consider the random process $(Y_m)_{m\ge1}$ on $\mathcal{Y} := \mathcal{K} \times \mathcal{A} \times \Delta(\mathcal{K})$ defined by $Y_m := (K_m, A_m, P_m)$. We claim that, under σ^*, the process $(Y_m)_{m\ge1}$ is a Markov chain. Indeed, given $m \ge 1$ and $(Y_1, \dots, Y_m) \in \mathcal{Y}^m$, Y_{m+1} is generated by the following procedure:

1. Draw a pair (K_{m+1}, S_m) according to $q(K_m, A_m)$.
2. Compute P_{m+1} using Bayes' rule according to P_m and S_m.
3. Draw the next action A_{m+1} according to $\sigma^*(P_{m+1})$.

By construction, the law of Y_{m+1} depends only on Y_m, and therefore $(Y_m)_{m\ge1}$ is a Markov chain.

Define $\nu^* \in \Delta(\mathcal{Y})$ by fixing the third marginal to μ^*, and for all $p \in \Delta(\mathcal{K})$, $\nu^*(\cdot \mid p) \in \Delta(\mathcal{K} \times \mathcal{A})$ is $p \otimes \sigma^*(p)$. We claim that ν^* is an invariant measure for $(Y_m)_{m\ge1}$. Indeed, fixing σ^* as the strategy for the player, if P_1 is drawn according to μ^*, then, because μ^* is σ^*-invariant, the third marginal of Y_m follows μ^*, for all $m \ge 1$. Moreover, conditional on P_m, the random variables K_m and A_m are independent: the conditional distribution of K_m is P_m, and the one of A_m is $\sigma^*(P_m)$. Thus, ν^* is an invariant measure of $(Y_m)_{m\ge1}$.

The strategy $\sigma^* : \Delta(\mathcal{K}) \to \Delta(\mathcal{A})$ is a (stationary) mixed strategy, and we are looking for a deterministic strategy $\sigma \in \Sigma$. To derandomize this strategy, note that σ^* starting from any $p \in \Delta(\mathcal{K})$ is strategically equivalent to a p-dependent element of $\Delta(\Sigma)$, that is, a distribution over pure (not necessarily stationary) strategies (Kuhn's theorem; see Feinberg [11]). To simplify notation, we still denote this equivalent strategy σ^* and omit its dependence in p.

Define $f : \mathcal{Y} \to [0,1]$ by $f(k,a,p) := g(k,a)$, a measurable function. Applying an ergodic theorem in Hernández-Lerma and Lasserre [15, theorem 2.5.1, p. 37], we know that $\exists f^*$ integrable with respect to μ^* such that for all $p \in \operatorname{Supp}(\mu^*)$ and $\sigma \in \operatorname{Supp}(\sigma^*)$,

$$\left(\frac{1}{n}\sum_{m=1}^{n} f(K_m, A_m, P_m) = \frac{1}{n}\sum_{m=1}^{n} G_m\right) \underset{[n\to\infty]}{\longrightarrow} f^*(K_1, a_1, p) \quad \mathbb{P}^p_\sigma\text{-a.s.},$$

where f^* satisfies that $\int_{\Delta(\mathcal{K})} f^*(y)\nu^*(dy) = \int_{\Delta(\mathcal{K})} f(y)\nu^*(dy)$.

We claim that for all $p \in \operatorname{Supp}(\mu^*)$ and $\sigma \in \operatorname{Supp}(\sigma^*)$, we have that, for all $k \in \operatorname{Supp}(p)$,

$$f^*(k, a_1, p) = \gamma^k_\infty(\sigma),$$

where a_1 is the first action according to σ (formally, $a_1 = \sigma(\emptyset) = \sigma^*(p)$).

Indeed, take $p \in \operatorname{Supp}(\mu^*)$, $\sigma \in \operatorname{Supp}(\sigma^*)$, and $k \in \operatorname{Supp}(p)$, and then

$$\gamma^k_\infty(\sigma) = \mathbb{E}^k_\sigma\left(\liminf_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} G_m\right) = \mathbb{E}^k_\sigma\big(f^*(K_1, a_1, p)\big) = f^*(k, a_1, p).$$

Because $\operatorname{Supp}(P^*) \subseteq \{k \in \mathcal{K} : \exists p \in \operatorname{Supp}(\mu^*) \text{ s.t. } k \in \operatorname{Supp}(p)\}$, property (A.2) is satisfied.

Let us now turn to property (A.3). We claim that, μ^*-*a.s.* and σ^*-a.s.,

$$\gamma^p_\infty(\sigma) = v_\infty(p).$$

Indeed, note that

$$\begin{aligned}
\int_{\Delta(\mathcal{K})}\int_{\Sigma} \gamma^p_\infty(\sigma)\sigma^*(d\sigma)\mu^*(dp) &= \int_{\Delta(\mathcal{K})} f^*(y)\nu^*(dy) && (\text{definitionof } \nu^*)\\
&= \int_{\Delta(\mathcal{K})} f(y)\nu^*(dy)\\
&= \int_{\Delta(\mathcal{K})} g(p, \sigma^*(p))\mu^*(dp) && (\text{definitionof } \nu^*)\\
&= \int_{\Delta(\mathcal{K})} v_\infty(p)\mu^*(dp) && (\text{Lemma A.8}).
\end{aligned}$$

By definition of v_∞, for all $\sigma \in \Sigma$, $\gamma^{\cdot}_\infty(\sigma) \le v_\infty(\cdot)$. Therefore, by positivity, we can conclude that, μ^*-a.s. and σ^*-a.s., $\gamma^p_\infty(\sigma) = v_\infty(p)$. It follows that

$$\gamma^{P^*}_\infty(\sigma) = v_\infty(P^*),$$

and property (A.3) is satisfied. □

References

[1] Arapostathis A, Borkar V, Fernández-Gaucherand E, Ghosh M, Marcus S (1993) Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control Optim.* 31(2):282–344.

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

[3] Bellman R (1957) A Markovian decision process. *J. Math. Mech.* 6(5):679–684.

[4] Blackwell D (1962) Discrete dynamic programming. *Ann. Math. Statist.* 33(2):719–726.

[5] Bonet B, Geffner H (2009) Solving POMDPs: RTDP-bel vs. point-based algorithms. *Proc. 21st Internat. Joint Conf. Artificial Intelligence* (Morgan Kaufmann, San Francisco), 1641–1646.

[6] Borkar V (2000) Average cost dynamic programming equations for controlled Markov chains with partial observations. *SIAM J. Control Optim.* 39(3):673–681.

[7] Bukharaev RG (1980) Probabilistic automata. *J. Math. Sci.* 13:359–386.

[8] Cerný P, Chatterjee K, Henzinger TA, Radhakrishna A, Singh R (2011) Quantitative synthesis for concurrent programs. Gopalakrishnan G, Qadeer S, eds. *Proc. Internat. Conf. Comput. Aided Verification*, Lecture Notes in Computer Science, vol. 6806 (Springer, Berlin), 243–259.

[9] Chatterjee K (2007) Concurrent games with tail objectives. *Theoret. Comput. Sci.* 388(1–3):181–198.

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

[11] Feinberg E (1996) On measurability and representation of strategic measures in Markov decision processes. Ferguson TS, Shapley LS, MacQueen JB, eds. *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell* (Institute of Mathematical Statistics), 29–43.

[12] Filar J, Vrieze K (1997) *Competitive Markov Decision Processes* (Springer, New York).

[13] Hansen KA, Ibsen-Jensen R, Neyman A (2018) Absorbing games with a clock and two bits of memory. Working paper, University of Glasgow, Scotland, UK.

[14] Hansen KA, Ibsen-Jensen R, Neyman A (2018) The big match with a clock and a bit of memory. *Proc. 2018 ACM Conf. Econom. Comput.* (ACM, New York), 149–150.

[15] Hernández-Lerma O, Lasserre JB (2003) Markov Chains and Invariant Probabilities (Birkhäuser, Basel, Switzerland).

[16] Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: A survey. *J. Artificial Intelligence Res*. 4:237–285.

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

[18] Neyman A, Sorin S (2010) Repeated games with public uncertain duration process. *Internat. J. Game Theory*. 39(1–2):29–52.

[19] Paz A (1971) *Introduction to Probabilistic Automata*. Computer Science and Applied Mathematics (Academic Press, Cambridge, MA).

[20] Rabin M (1963) Probabilistic automata. *Inform. Control* 6(3):230–245.

[21] Renault J (2011) Uniform value in dynamic programming. *J. Eur. Math. Soc.* 13(2):309–330.

[22] Renault J, Venel X (2016) Long-term values in Markov decision processes and repeated games, and a new distance for probability spaces. *Math. Oper. Res.* 42(2):349–376.

[23] Rosenberg D, Solan E, Vieille N (2002) Blackwell optimality in Markov decision processes with partial observation. *Ann. Statist.* 30(4):1178–1193.

[24] Shapley L (1953) Stochastic games. *Proc. Natl. Acad. Sci. USA*. 39(10):1095–1100.

[25] Solan E (2003) Continuity of the value of competitive Markov decision processes. *J. Theoret. Probab.* 16(4):831–845.

[26] Solan E, Vieille N (2010) Computing uniformly optimal strategies in two-player stochastic games. *Econom. Theory* 42(1):237–253.

[27] Venel X, Ziliotto B (2016) Strong uniform value in gambling houses and partially observable Markov decision processes. *SIAM J. Control Optim.* 54(4):1983–2008.

[28] Venel X, Ziliotto B (2021) History-dependent evaluations in POMDPs. *SIAM J. Control Optim.* Forthcoming.