The Annals of Statistics
2002, Vol. 30, No. 4, 1178–1193

BLACKWELL OPTIMALITY IN MARKOV DECISION PROCESSES WITH PARTIAL OBSERVATION

BY DINAH ROSENBERG, EILON SOLAN AND NICOLAS VIEILLE

Université Paris Nord, Northwestern University and Tel Aviv University and HEC

A Blackwell ε-optimal strategy in a Markov Decision Process is a strategy that is ε-optimal for every discount factor sufficiently close to 1.

We prove the existence of Blackwell ε-optimal strategies in finite Markov Decision Processes with partial observation.

1. Introduction. A well-known result by Blackwell [3] states that, in any Markov Decision Process (MDP hereafter) with finitely many states and finitely many actions, there is a pure stationary strategy that is optimal, for *every* discount factor close enough to one. This strong optimality property is now referred to as Blackwell optimality.

In this paper we study finite MDPs with partial observations (p.o. hereafter); that is, finite MDPs in which at the end of every stage, the decision maker receives a signal that depends randomly on the current state and on the action that has been chosen, but he observes neither the state nor his daily payoff (see, e.g., [2] and the references in [7]). MDPs with p.o. arise naturally in many contexts, such as models of machine replacement and quality control problems (see [12] and the references therein for this and additional applications), telecommunication networks (see [1] and the references therein), and intra-seasonal decisions of fishing vessel operators [9].

Here we address the problem of existence of Blackwell optimal strategies for a finite MDP with p.o. We prove that, in any such MDP and for every ε, there is a strategy that is Blackwell ε-optimal; that is, ε-optimal for *every* discount factor close enough to one. The strategy we construct is moreover ε-optimal in the n-stage MDP, for every n large enough. We also provide an example where there is no Blackwell zero-optimal strategy.

The standard approach to an MDP with p.o. is to convert it into an auxiliary MDP with full observation and Borel state space. The conditional distribution over the state space Ω given the available information (sequence of past signals and past actions) plays the role of the state variable in the auxiliary MDP. This approach has been developed for instance in [14], [15] and [17]. An alternative state variable is defined in [5]. One then looks for optimal stationary strategies (strategies such that the action chosen in any given stage is only a function of the belief held on

Received May 2000; revised December 2001.

AMS 2000 subject classifications. 60J20, 62C99.

Key words and phrases. Blackwell optimality, MDP, partial observation.

the underlying state in Ω). A commonly used criterion is the long-run average cost criterion; see, for example, [4] and [6].

If the sets of actions and signals are finite, then the conditional distribution at every stage can take only finitely many values. In particular, the auxiliary MDP with full observation is defined over a countable state space. If, in addition, there are no signals, then the transitions in the auxiliary MDP are deterministic. Lehrer and Sorin [10] provided an example of an MDP with countable state space, finitely many actions and deterministic transitions where the limit of the discounted values (as the discount factor goes to 1) and the limit of the n-stage process (as n goes to ∞) exist, differ, and differ from the value under the long-run average cost criterion. It follows that in their example there is no Blackwell ε-optimal strategy for any $\varepsilon > 0$ sufficiently small. It is well known that if the state space is countable and the limit of the discounted values as well as the limit of the n-stage process exist and are equal, there need not be a Blackwell optimal strategy (Example 7.1.3 in [16]). Moreover, even if a Blackwell ε-optimal strategy exists, it need not be stationary (Example 7.1.5 in [16]).

To guarantee the existence of optimal strategies in MDPs with Borel state space one has to impose further assumptions on the transitions. These assumptions usually have the flavor of an irreducibility condition. For MDPs that arise from an MDP with p.o., these conditions may be difficult to interpret in terms of the underlying data; see, for instance, [2], page 329, [13], page 415 or [16], page 158.

In the present paper we do not follow this approach but rather use the structure on the auxiliary MDP that is derived from the underlying MDP. Specifically, using a sequence of optimal strategies in the n-stage MDP, and using the compactness of the state space of the auxiliary MDP and the continuity of the payoff on this space, we construct a Blackwell ε-optimal strategy.

In Section 2, we present the model and the main results. In Section 3, we show on an example that the result is in some respect tight. In Section 6, we construct a Blackwell ε-optimal strategy. This strategy is neither pure nor stationary. In the case of degenerate observation (the decision maker receives no information whatsoever), we construct a pure, stationary Blackwell ε-optimal strategy. Part of this proof serves as an introduction for the general case. It is therefore presented in Section 5. Section 4 contains a number of preliminary results that are used in both proofs.

2. The model and the main results.

Given a set M, we denote by $\Delta(M)$ the set of probability distributions over M, and we identify M with the set of extreme points of $\Delta(M)$.

A *Markov decision process with partial observation* is given by: (i) a state space Ω, (ii) an action set A, (iii) a signal set S, (iv) a transition rule $q : \Omega \times A \to \Delta(S \times \Omega)$, (v) a payoff function $r : \Omega \times A \to \mathbf{R}$ and (vi) a probability distribution $x_1 \in \Delta(\Omega)$.

We assume that Ω, A and S are finite sets. Extensions to more general cases are discussed below. Without loss of generality, we assume that $0 \leq r(\omega, a) \leq 1$ for every $(\omega, a) \in \Omega \times A$.

The description of the model [i.e., (i)–(vi)] is known to the decision maker.

An initial state ω_1 is drawn according to x_1. At every stage n the decision maker chooses an action $a_n \in A$, and a pair $(s_n, \omega_{n+1}) \in S \times \Omega$ of a signal and a new state is drawn according to $q(\omega_n, a_n)$. The decision maker is informed of the signal s_n, but not of the new state ω_{n+1} or the payoff $r(\omega_n, a_n)$.

Thus, the information available to the decision maker at stage n is the finite sequence $a_1, s_1, a_2, s_2, \dots, a_{n-1}, s_{n-1}$ and a (*behavioral*) *strategy* for the decision maker is a function that assigns for every such sequence a probability distribution over $\Delta(A)$. We set $H_n = (A \times S)^{n-1}$, and we denote respectively by $H = \bigcup_{n \geq 1} H_n$ and $H_\infty = (A \times \Omega \times S)^{\mathbf{N}}$ the set of finite histories and infinite plays. We denote by $\mathcal{H}_n$ the algebra of cylinder sets over H_∞ induced by H_n.

Each strategy σ, together with the initial distribution x_1, induces a probability distribution $\mathbf{P}_{x_1,\sigma}$ over $(H_\infty, \mathcal{H}_\infty)$, where $\mathcal{H}_\infty = \sigma(\mathcal{H}_n, n \geq 1)$. Expectations under $\mathbf{P}_{x_1,\sigma}$ are denoted by $\mathbf{E}_{x_1,\sigma}$. All norms in the paper are supremum norms, unless otherwise specified.

We let

$$\gamma_n(x_1, \sigma) = \mathbf{E}_{x_1,\sigma}\big[\big(r(\omega_1, a_1) + \cdots + r(\omega_n, a_n)\big)/n\big]$$

denote the expected average payoff in the first n stages.

We denote by $v_n(x_1) = \sup_\sigma \gamma_n(x_1, \sigma)$ the value of the n-stage problem.

For every $\lambda \in (0, 1)$ and every strategy σ we define the λ-discounted payoff as

$$\gamma_\lambda(x_1, \sigma) = \mathbf{E}_{x_1,\sigma}\left[(1 - \lambda) \sum_{m=1}^{\infty} \lambda^{m-1} r(\omega_m, a_m)\right]$$

and the discounted value by

$$v_\lambda(x_1) = \sup_\sigma \gamma_\lambda(x_1, \sigma).$$

DEFINITION 1. An MDP with p.o. is said to have a *value* (w.r.t. a given initial distribution x_1) if both limits $\lim_{n\to\infty} v_n(x_1)$ and $\lim_{\lambda\to 1} v_\lambda(x_1)$ exist and are equal.

If the MDP has a value, we denote it by $v(x_1) = \lim_{n\to\infty} v_n(x_1) = \lim_{\lambda\to 1} v_\lambda(x_1)$.

DEFINITION 2. For a given initial distribution x_1 and $\varepsilon \geq 0$, a strategy σ is (*Blackwell*) ε-*optimal* (w.r.t. x_1) if there exist a positive integer $N_0 \in \mathbf{N}$, and $\lambda_0 \in (0, 1)$ such that:

$$\gamma_n(x_1, \sigma) \geq v_n(x_1) - \varepsilon, \qquad \forall\, n \geq N_0 \tag{1}$$

and

$$\gamma_\lambda(x_1, \sigma) \geq v_\lambda(x_1) - \varepsilon, \qquad \forall\, \lambda \in (\lambda_0, 1). \tag{2}$$

Our first main result is that the value always exists, as well as ε-optimal strategies.

THEOREM 1. *If* Ω, *A and S are finite*, *then the MDP with p.o. has a value* $v(x_1)$ *w.r.t. every initial distribution* $x_1 \in \Delta(\Omega)$. *Moreover*, *for every* $\varepsilon > 0$ *and every* $x_1 \in \Delta(\Omega)$, *there is a* (*Blackwell*) ε*-optimal strategy w.r.t.* x_1.

In the case where $|S| = 1$, that is, the decision maker receives no informative signal, we get a stronger result.

To state this result we need additional notions. For $n \geq 1$, we denote by y_n the conditional distribution of ω_n given $\mathcal{H}_n$: for each $\omega \in \Omega$, $y_n[\omega]$ is the posterior probability in stage n that the process is at state ω given the information available to the decision maker (we do not assume here that $|S| = 1$). Thus, $y_1 = x_1$. Observe that the value $y_n(h_n) \in \Delta(\Omega)$ of y_n after a given history h_n may be computed without knowledge of the strategy. y_n is therefore a function $H_n \to \Delta(\Omega)$ or, equivalently, a random variable $(H_\infty, \mathcal{H}_n) \to \Delta(\Omega)$. Clearly, the *distribution* of y_n is influenced by the strategy that is followed.

A *pure strategy* is a strategy $\sigma : H \to \Delta(A)$, such that $\sigma(h) \in A$ for each $h \in H$. A strategy is *stationary* if $\sigma(h_n)$ depends only on $y_n(h_n)$, the posterior probability at stage n.

Our second result is that if $|S| = 1$, the ε-optimal strategies can be chosen to be pure and stationary.

THEOREM 2. *If* Ω *and A are finite*, *and* $|S| = 1$, *then for every* $\varepsilon > 0$ *and every* $x_1 \in \Delta(\Omega)$ *there exists a pure stationary* (*Blackwell*) ε*-optimal strategy w.r.t.* x_1.

COMMENT 1. We claim here that to prove Theorems 1 and 2, it is enough to prove for all $x_1 \in \Delta(\Omega)$ that $v(x_1) = \lim_{n\to\infty} v_n(x_1)$ exists and (1) holds. Since Ω is finite [so that $\Delta(\Omega)$ is compact], Proposition 1 below implies that $(v_n(x_1))$ converges to $v(x_1)$ uniformly in $x_1 \in \Delta(\Omega)$. It follows by [10] that $\lim_{\lambda\to 1} v_\lambda(x_1)$ exists and is equal to $\lim_{n\to\infty} v_n(x_1)$. Moreover, by Eq. (1) of [10], it follows that $\liminf_{\lambda\to 1} \gamma_\lambda(x_1, \sigma) \geq \liminf_{n\to\infty} \gamma_n(x_1, \sigma)$. Hence (2) holds as well.

COMMENT 2. A probability distribution over pure strategies is called a *mixed strategy*. An MDP with p.o. can be described as a single player game in extensive form with perfect recall, so that Kuhn's theorem [8] applies. In particular, for every behavioral strategy σ there exists a mixed strategy π such that σ and π induce the same probability distribution over H_∞; that is, the probability distribution over H_∞ obtained by first choosing a pure strategy f according to π, and then following f, coincides with $\mathbf{P}_{x_1,\sigma}$.

By Theorem 1 an ε-optimal strategy exists for every $\varepsilon > 0$. Nevertheless, it is not clear that a *pure* ε-optimal strategy exists as well. Indeed, by Kuhn's theorem, for every fixed $n \geq 1$ there exists a strategy f_n in the support of π, such that $\gamma_n(x_1, f_n) \geq \gamma_n(x_1, \sigma)$. However, it is not clear at all that f_n can be chosen *independently* of n.

An implication of Kuhn's theorem is that if there is no pure zero-optimal strategy then there is no behavioral zero-optimal strategy. Indeed, fix a behavioral strategy σ, and let π be the equivalent mixed strategy. Since no pure strategy f is zero-optimal, for every fixed pure strategy f there are infinitely many n's such that $\gamma_n(x_1, f) < v_n(x_1)$. It follows that $\gamma_n(x_1, \sigma) < v_n(x_1)$ for infinitely many n's as well.

OPEN PROBLEMS. There are several natural questions that arise. First, does there exist a single strategy that is Blackwell ε-optimal for *every* $\varepsilon > 0$? Second, do pure stationary Blackwell ε-optimal strategies exist in general?

It is also desirable to understand the evolution of the posterior distribution under ε-optimal strategies. Simple examples show that even in the case of no signals the sequence of posterior distributions need not be stationary, eventually stationary, or converge to some limit. We do not know whether there exists an ε-optimal strategy such that this sequence is asymptotically periodic. This point is related to the turnpike property that is studied in various economic models (see [11] and the references therein).

3. An example. Define an MDP with no signals as follows. Set $\Omega = \{*, \omega\}$, and $A = \{a^1, a^2\}$. The transition rule q is given by

$$q(*|*, a) = 1 \qquad \text{for each } a \in A,$$
$$q(\omega|\omega, a^1) = 1, \qquad q(\omega|\omega, a^2) = \tfrac{1}{2}.$$

The payoff function r is given by

$$r(*, a^1) = 1, \quad r(*, a^2) = 0 \quad \text{and} \quad r(\omega, a) = 0 \qquad \text{for } a \in A.$$

We identify a probability distribution over Ω with the probability assigned to ω. The MDP starts from state ω, so that $x_1 = 1$. Since there are no signals, the probability y_n that the process is in state ω at stage n is determined by past actions, and a pure strategy may be identified with a sequence $(a_n)_{n \in \mathbf{N}}$ of actions.

State $*$ is absorbing. Therefore, $a_n = a^1$ implies that $\omega_{n+1} = \omega_n$, hence $y_{n+1} = y_n$, while $a_n = a^2$ implies that $y_{n+1} = y_n/2$.

The value of this MDP is equal to one. Indeed, given $\varepsilon > 0$, let σ be the (stationary) strategy that plays a^2 in the first $N = \lfloor -\log_2 \varepsilon \rfloor + 2$ stages, and plays a^1 afterwards. Given σ, one has $y_{N+1} < \varepsilon$. Therefore, $\mathbf{E}_{x_1,\sigma}[r(\omega_n, a_n)] = 1 - y_{N+1} > 1 - \varepsilon$ for each $n > N$. In particular, $\liminf_{n\to\infty} \gamma_n(x_1, \sigma) = \liminf_{\lambda\to 1} \gamma_\lambda(x_1, \sigma) > 1 - \varepsilon$. Since $v_n(x_1) \leq 1$, and $v_\lambda(x_1) \leq 1$, the value is indeed equal to 1. In particular, $\lim_{\lambda\to 1} v_\lambda(x_1) = \lim_{n\to\infty} v_n(x_1) = 1$.

We now claim that in this example there is no zero-optimal strategy. It suffices to prove that there is no pure zero-optimal strategy. Let $\sigma = (a_n)_{n\in\mathbf{N}}$ be a pure strategy. We distinguish three (nonexclusive) cases.

Case 1. There exists $N \in \mathbf{N}$, such that $a_n = a^1$ for every $n \geq N$.

In that case, the sequence (y_n) is constant from stage N on. Therefore, $\lim_{n\to\infty} \gamma_n(x_1, \sigma) = \lim_{\lambda\to 1} \gamma_\lambda(x_1, \sigma) = 1 - y_N < 1$. In particular, $\gamma_\lambda(x_1, \sigma) < v_\lambda(x_1)$ for λ close to one, and therefore σ is not zero-optimal.

Case 2. There exists $N \in \mathbf{N}$, such that $a_n = a^2$ for every $n \geq N$.

In that case, $\mathbf{E}_\sigma[r(\omega_n, a_n)] = 0$ for each $n \geq N$. Therefore, $\lim_{n\to\infty} \gamma_n(x_1, \sigma) = \lim_{\lambda\to 1} \gamma_\lambda(x_1, \sigma) = 0$, and therefore σ is not zero-optimal.

Case 3. There exists $n_0 \in \mathbf{N}$, such that $a_{n_0} = a^1$ and $a_{n_0+1} = a^2$. Denote by τ the strategy obtained from σ by permutation of a_{n_0} and a_{n_0+1}. Observe that

$$\begin{aligned}\mathbf{E}_\tau[r(\omega_n, a_n)] &= \mathbf{E}_\sigma[r(\omega_n, a_n)] \qquad \text{for each } n \in \mathbf{N}\backslash\{n_0, n_0+1\},\\ \mathbf{E}_\tau[r(\omega_{n_0}, a_{n_0})] &= \mathbf{E}_\sigma[r(\omega_{n_0+1}, a_{n_0+1})] = 0,\\ \mathbf{E}_\tau[r(\omega_{n_0+1}, a_{n_0+1})] &> \mathbf{E}_\sigma[r(\omega_{n_0}, a_{n_0})].\end{aligned}$$

Therefore, $\gamma_\lambda(x_1, \tau) > \gamma_\lambda(x_1, \sigma)$ for λ close to one. In particular, σ is not zero-optimal for λ close to one.

A natural question arises. Does there exist a strategy that is ε-optimal for each $\varepsilon > 0$? We claim that in this example there is such a (nonstationary) pure strategy. Indeed, let $\sigma = (a_n)_{n\in\mathbf{N}}$ be a pure stationary strategy. Since $y_{n+1} = y_n$ whenever $a_n = a^1$, the stationarity of σ implies that $a_{n+1} = a^1$ as soon as $a_n = a^1$. This implies that the sequence (a_n) is eventually constant, that is, it must be that either case 1 or case 2 above holds. In both cases, σ fails to be ε-optimal, provided ε is small enough.

Let now $\sigma = (a_n)$ be any sequence such that the subset $N^2 = \{n \in \mathbf{N}, a_n = a^2\}$ of $\mathbf{N}$ is infinite and has density zero. Since N^2 is infinite, the sequence (y_n) converges to zero under σ. Therefore,

$$\lim_{n\to\infty, n\notin N^2} \mathbf{E}_\sigma[r(\omega_n, a_n)] = 1. \tag{3}$$

Since N^2 has density zero, (3) yields $\lim_{n\to\infty} \gamma_n(x_1, \sigma) = \lim_{\lambda\to 1} \gamma_\lambda(x_1, \sigma) = 1$.

As an anonymous referee mentioned, there is also a stationary nonpure strategy that is ε-optimal for every $\varepsilon > 0$: if $1/(n+1) < y_n \leq 1/n$ play a^2 with probability $1/n$.

4. Preliminaries.

The purpose of this section is to introduce several general results. The first result is standard. It asserts that, given $N \in \mathbf{N}$, there exists a pure optimal strategy in the N-stage MDP such that the action played at stage n depends only on n and y_n.

LEMMA 1. *For each $N \geq 1$, there exists a pure strategy σ_N such that $\gamma_N(x_1, \sigma_N) = v_N(x_1)$ and $\sigma_N(h_n)$ is only a function of n and $y_n(h_n)$.*

The lemma follows from standard dynamic programming arguments, hence its proof is omitted.

Whenever in the sequel we refer to optimal strategies in the n-stage problem, we mean a pure strategy that satisfies the condition in Lemma 1.

Given $m < n$, we denote by

$$\gamma_{m,n}(x_1,\sigma) = \mathbf{E}_{x_1,\sigma}\left[\frac{1}{n-m+1}\big(r(\omega_m,a_m)+\cdots+r(\omega_n,a_n)\big)\right]$$

the expected average payoff from stage m up to stage n. Thus, $\gamma_n(x_1,\sigma) = \gamma_{1,n}(x_1,\sigma)$.

PROPOSITION 1. *Let $x, x' \in \Delta(\Omega)$. For every strategy σ and every $m < n$,*

$$|\gamma_{m,n}(x,\sigma) - \gamma_{m,n}(x',\sigma)| \le \|x - x'\|_1,$$

where $\|\cdot\|_1$ is the L_1-norm.

PROOF. Let $n \ge 1$ and $\overline{h}_n \in H_n$ be given. Observe that, for every $x \in \Delta(\Omega)$ and for every strategy σ, one has

$$\mathbf{P}_{x,\sigma}(h_n = \overline{h}_n) = \sum_{\omega\in\Omega} x(\omega)\mathbf{P}_{\omega,\sigma}(h_n = \overline{h}_n).$$

In particular, $\mathbf{E}_{x,\sigma}[r(s_n,a_n)] = \sum_{\omega\in\Omega} x(\omega)\mathbf{E}_{\omega,\sigma}[r(s_n,a_n)]$. Since $0 \le r \le 1$, the result follows. □

PROPOSITION 2. *Let a strategy σ, $\varepsilon \in (0,1)$ and $n \in \mathbf{N}$ be given, and set*

$$N = \inf\{k \in \mathbf{N},\ \text{s.t. } \gamma_m(x_1,\sigma) \ge \gamma_n(x_1,\sigma) - \varepsilon \text{ for every } k \le m \le n\}. \tag{4}$$

Then $N \le 1 + (1-\varepsilon)n$. Moreover,

$$\gamma_{N,m}(x_1,\sigma) \ge \gamma_n(x_1,\sigma) - \varepsilon \qquad \text{for every } N \le m \le n. \tag{5}$$

Given $\varepsilon > 0$ and σ, let $N_n = N_n(\varepsilon,\sigma)$ denote the integer associated with n in (4). Note that $\liminf_{n\to\infty}(n - N_n) = +\infty$. Proposition 2 has the same flavor as Proposition 2 in [10].

PROOF OF PROPOSITION 2. Clearly, $N \le n$. Note that if $N > 1$ then $\gamma_{N-1}(x_1,\sigma) < \gamma_n(x_1,\sigma) - \varepsilon$.

We first show that $N \le 1 + (1-\varepsilon)n$. Indeed, otherwise, $N > 1$, hence $\gamma_{N-1}(x_1,\sigma) < \gamma_n(x_1,\sigma) - \varepsilon$. Since $0 \le r \le 1$,

$$\gamma_n(x_1,\sigma) \le \frac{N-1}{n}\gamma_{N-1}(x_1,\sigma) + \frac{n-N+1}{n} < \gamma_n(x_1,\sigma) - \varepsilon + \varepsilon = \gamma_n(x_1,\sigma),$$

a contradiction.

Next we show that (5) holds. Fix an integer m such that $N \le m \le n$. If $N = 1$, by the definition of N we get $\gamma_{N,m}(x_1,\sigma) = \gamma_m(x_1,\sigma) \ge \gamma_n(x_1,\sigma) - \varepsilon$. If $N > 1$, $\gamma_{N-1}(x_1,\sigma) < \gamma_n(x_1,\sigma) - \varepsilon$, while $\gamma_m(x_1,\sigma) \ge \gamma_n(x_1,\sigma) - \varepsilon$. It follows that $\gamma_{N,m}(x_1,\sigma) > \gamma_n(x_1,\sigma) - \varepsilon$. □

5. The case of "no signals." This section is devoted to the proof of Theorem 2. Thus, we assume that no signal is available. The initial distribution x_1 is fixed throughout the section.

In this case, a pure strategy is reduced to a sequence of actions: the action that is played at each stage. Moreover, if σ is pure, the posterior distribution at stage n depends deterministically on σ. We write $y_n(\sigma)$ for the posterior distribution at stage n:

$$y_n(\sigma)[\omega] = \mathbf{P}_{x_1,\sigma}(\omega_n = \omega).$$

If $\sigma = (a_1, a_2, \dots) \in A^{\mathbf{N}}$ is a strategy, we define for every positive integer $m \in \mathbf{N}$ the truncated strategy $\sigma^m = (a_m, a_{m+1}, \dots)$ and the prefix ${}^m\sigma = (a_1, \dots, a_m)$.

Define $w = \limsup_{n\to\infty} v_n(x_1)$, and fix $\varepsilon \in (0, 1)$. Let $(n_i)_{i\in\mathbf{N}}$ be an increasing subsequence such that $\lim_{i\to\infty} v_{n_i}(x_1) = w$ and $|v_{n_i}(x_1) - w| < \varepsilon$ for every $i \in \mathbf{N}$. Let σ_i be a pure optimal strategy in the n_i-stage problem (that satisfies the condition of Lemma 1). Thus, $\gamma_{n_i}(x_1, \sigma_i) = v_{n_i}(x_1)$.

Given $i \in \mathbf{N}$, we let $N_i \le 1 + (1 - \varepsilon)n_i$ be the integer obtained by applying Proposition 2 to σ_i, ε and n_i.

For notational simplicity, we set $y_i = y_{N_i}(\sigma_i)$. Since Ω is finite, $\Delta(\Omega)$ is compact, hence there exists $y \in \Delta(\Omega)$ and a subsequence of $\{y_i\}$, still denoted by $\{y_i\}$, such that $\|y_i - y\| < \varepsilon/2$, for each $i \in \mathbf{N}$. In particular, $\|y_i - y_1\| < \varepsilon$ for every $i \in \mathbf{N}$.

For each $i \in \mathbf{N}$ define π_i as: follow σ_1 up to N_1, switch to $\sigma_i^{N_i}$ at stage N_1. Formally,

$$\pi_i(n) = \begin{cases} \sigma_1(n), & \text{for } 1 \le n \le N_1 - 1, \\ \sigma_i(N_i + n - N_1), & \text{for } N_1 \le n. \end{cases}$$

Set $m_i = N_1 + n_i - N_{_i}$. Note that $\liminf_{i\to\infty} m_i = +\infty$.

PROPOSITION 3. *Let $i \in \mathbf{N}$ be sufficiently large so that $(N_1 - 1)/\varepsilon < m_i$. If m satisfies $(N_1 - 1)/\varepsilon < m \le m_i$ then*

$$\gamma_m(x_1, \pi_i) \ge w - 4\varepsilon.$$

Proposition 3 asserts that each π_i gives high payoff in *all* m-stage problems, provided m is sufficiently large (but smaller than m_i). Moreover, the lower bound on m is independent of i.

PROOF OF PROPOSITION 3. Fix an integer m such that $(N_1 - 1)/\varepsilon < m \le m_i$. By construction, $y_{N_1}(\pi_i) = y_1$, hence

$$\begin{aligned}\gamma_m(x_1, \pi_i) &= \frac{N_1 - 1}{m}\gamma_{N_1-1}(x_1, \pi_i) + \frac{m - N_1 + 1}{m}\gamma_{N_1,m}(x_1, \pi_i) \\ &= \frac{N_1 - 1}{m}\gamma_{N_1-1}(x_1, \pi_i) + \frac{m - N_1 + 1}{m}\gamma_{m-N_1+1}(y_1, \pi_i^{N_1}).\end{aligned}$$

By the assumption on m, $(m - N_1 + 1)/m > 1 - \varepsilon$. Since $\|y_1 - y_i\| < \varepsilon$, we get by Proposition 1, and since payoffs are nonnegative,

$$\gamma_m(x_1, \pi_i) \geq (1-\varepsilon)\bigl(\gamma_{m-N_1+1}(y_i, \pi_i^{N_1}) - \varepsilon\bigr) = (1-\varepsilon)\bigl(\gamma_{N_i, m-N_1+N_i}(x_1, \sigma_i) - \varepsilon\bigr).$$

Since $m - N_1 + N_i > N_i$, Proposition 2 implies that $\gamma_{N_i, m-N_1+N_i}(x_1, \sigma_i) \geq v_{n_i} - \varepsilon > w - 2\varepsilon$. One then has

$$\gamma_m(x_1, \pi_i) > (1-\varepsilon)(w - 3\varepsilon) > w - 4\varepsilon,$$

as desired. □

PROPOSITION 4. *In the case* $|S| = 1$, *the MDP with p.o. has a value* $v(x_1)$ *w.r.t. every initial distribution* $x_1 \in \Delta(\Omega)$.

PROOF. Since A is finite, by a diagonal extraction argument there exists a pure strategy π such that every prefix of π is a prefix of infinitely many π_i's: for each m, ${}^m\pi = {}^m\pi_i$ for infinitely many i. In particular, for every $m > N_1/\varepsilon$, $\gamma_m(x_1, \pi) > w - 4\varepsilon$. It follows that $v_m(x_1) > w - 4\varepsilon$. Since $\varepsilon > 0$ is arbitrary, one has $w = \lim_{n\to\infty} v_n(x_1)$ and π is a 4ε-optimal strategy. □

PROOF OF THEOREM 2. Let $\pi = (a_1, a_2, \dots)$ be a pure ε-optimal strategy; that is, there exists $n_0 \in \mathbf{N}$ such that $\gamma_n(x_1, \pi) \geq w - \varepsilon$ for every $n \geq n_0$. Let $y_n = y_n(\pi)$ be the posterior distribution at stage n.

Case 1. There exist $n_1 \in \mathbf{N}$ and $d \in \mathbf{N}$ such that $a_n = a_{n+d}$ and $y_n = y_{n+d}$ for every $n \geq n_1$.

Since π is ε-optimal, it follows that the expected average payoff along the period is at least $w - \varepsilon$:

$$\gamma_{n_1, n_1+d-1}(x_1, \pi) \geq w - \varepsilon.$$

We call a pure strategy π' *eventually stationary* if there exists $n_2 \in \mathbf{N}$ such that for every $n, m \geq n_2$,

$$y_n(\pi) = y_m(\pi) \Rightarrow a_n = a_m.$$

We now show by induction over the length d of the period of (a_n) that there exists a pure eventually stationary ε-optimal strategy π'.

If $d = 1$ then π is eventually stationary. If $d > 1$ and for every i, j that satisfy $n_1 \leq i < j < n_1 + d$ we have $y_i \neq y_j$ then π is eventually stationary as well. So assume that $d > 1$ and that $y_i = y_j$ for some i, j that satisfy $n_1 \leq i < j < n_1 + d$. If $\gamma_{i, j-1}(x_1, \pi) > w - \varepsilon$, the strategy $\pi' = (a_1, a_2, \dots, a_{n_1}, a_{n_1+1}, \dots, a_i, a_{i+1}, \dots, a_{j-1}, a_i, a_{i+1}, \dots, a_{j-1}, \dots)$ is 2ε-optimal and eventually periodic, with period $j - i < d$. If, on the other hand, $\gamma_{i, j-1}(x_1, \pi) \leq w - \varepsilon$, the strategy $\pi' = (a_1, a_2, \dots, a_{n_1}, a_{n_1+1}, \dots, a_{i-1}, a_j, a_{j+1}, \dots, a_{n_1+d-1},$

$a_{n_1}, a_{n_1+1}, \ldots, a_{i-1}, a_j, a_{j+1}, \ldots, a_{n_1+d-1}, \ldots)$ is ε-optimal and eventually periodic, with period $d-(j-i)<d$. In both cases, the induction hypothesis shows that the claim holds.

Thus, we assume w.l.o.g. that π is eventually stationary. In particular, there are $n_2 \geq n_1$ and $d \in \mathbf{N}$ such that $y_{n+d}=y_n$ for every $n \geq n_2$, and $y_n=y_m$ implies that $a_n=a_m$ for every $n, m \geq n_2$. Let $Y=\{y_n, n=1, \ldots, n_2+d-1\}$ be the set of all posterior distributions in the first n_2+d-1 stages. Consider the directed graph whose vertices are the elements in Y, and which contains the edge $(y, y') \in Y \times Y$ if and only if $(y, y')=(y_n, y_{n+1})$ for some $n \in \{1, \ldots, n_2+d-1\}$. Thus we connect with an edge any two consecutive elements in the finite sequence $(y_n)_{n=1}^{n_2+d}$.

Clearly there is a path from y_1 to any $y \in Y$. Let $y_1=y_{i_1}, y_{i_2}, \ldots, y_{i_k}$ be a shortest path that connects y_1 to the set $\{y_{n_2}, y_{n_2+1}, \ldots, y_{n_2+d-1}\}$. In particular, $y_{i_j} \neq y_{i_{j'}}$ for every $1 \leq j < j' \leq k$. Assume w.l.o.g. that $y_{i_k}=y_{n_2}$. Define

$$\pi''=(a_{i_1}, a_{i_2}, \ldots, a_{i_k-1}, a_{n_2}, a_{n_2+1}, \ldots, a_{n_2+d-1}, a_{n_2}, a_{n_2+1}, \ldots, a_{n_2+d-1}, \ldots).$$

By construction, $y_n(\pi'')=y_{i_n}(\pi)$ for each $n<k$, $y_k(\pi'')=y_{n_2}(\pi)$, and the sequence $(y_n(\pi''))_{n \geq k}$ coincides with the periodic sequence $(y_{n_2}(\pi), \ldots, y_{n_2+d-1}(\pi), y_{n_2}(\pi), \ldots, y_{n_2+d-1}(\pi), \ldots)$. Each of the posteriors $y_n(\pi'')$, $n < k+d$ appears only once, hence π'' is stationary. Since $\gamma_{n_2,n_2+d-1}(x_1, \pi) \geq w-\varepsilon$, we have $\gamma_n(x_1, \pi'') \geq w-2\varepsilon$ for every $n \geq k(n_2+d)/\varepsilon$.

Case 2. There are two integers $0<n_1<n_2$ such that $y_{n_1}=y_{n_2}$, and $\gamma_{n_1,n_2-1}(x_1, \pi) \geq w-\varepsilon$.

Define the strategy $\pi'=(a_1, a_2, \ldots, a_{n_1}, a_{n_1+1}, \ldots, a_{n_2-1}, a_{n_1}, \ldots, a_{n_2-1}, \ldots)$. Then π' is 2ε-optimal, and $(y_n(\pi'))$ is eventually periodic. We can then apply Case 1 to π'.

Case 3. There is some $y \in \Delta(\Omega)$ that appears infinitely often in the sequence $(y_n)_{n \in \mathbf{N}}$.

Since for every n sufficiently large, $\gamma_n(x_1, \pi) \geq w-\varepsilon$, it follows that there exist $n_1<n_2$ such that $y_{n_1}=y_{n_2}=y$ and $\gamma_{n_1,n_2-1}(x_1, \pi) \geq w-\varepsilon$. Apply now Case 2.

Case 4. None of the above holds.

Since Case 3 does not hold, every $y \in \Delta(\Omega)$ that appears in the sequence $(y_n)_{n \in \mathbf{N}}$, does so only finitely many times. Since Case 2 does not hold, the expected average payoff between two appearances of any $y \in \Delta(\Omega)$ in (y_n) is below $w-\varepsilon$.

Define a sequence $(i_k)_{k \in \mathbf{N}}$ as follows:

$$i_1=\max\{n \geq 1, y_n=y_1\}$$

and

$$i_{k+1}=\max\{n \geq 1, y_n=y_{i_k+1}\}. \tag{6}$$

In words, i_1 is the last occurrence of the initial distribution, i_2 is the last occurrence of the distribution at stage i_1+1, and so on. Since y_{i_k} appears only finitely many times in the sequence (y_n), the maximum in (6) is finite. Clearly $i_{k+1} > i_k$. Note that $y_{i_{k+1}} = y_{i_k+1}$, for each k.

Define now a strategy $\pi' = (a_{i_1}, a_{i_2}, a_{i_3}, \ldots)$. Since $y_{i_{k+1}} = y_{i_k+1}$, it follows by induction that

$$y_{i_{k+1}} = y(a_{i_1}, a_{i_2}, \ldots, a_{i_k}),$$

where $y(a_{i_1}, a_{i_2}, \ldots, a_{i_k})$ is the posterior probability held after playing actions $a_{i_1}, a_{i_2}, \ldots, a_{i_k}$. It also follows that no element in the sequence (y_{i_k}) appears twice. In particular, the strategy π' is stationary.

Recall that $\gamma_n(x_1, \pi) \geq w - \varepsilon$ for every $n \geq n_0$. We now argue that for every $k_0 \geq n_0$, $\gamma_{k_0}(x_1, \pi') \geq w - \varepsilon$. Set $n = i_{k_0}$ and $i_0 = 0$. Note that

$$n = \sum_{k=1}^{k_0} (i_k - i_{k-1}) = k_0 + \sum_{k \leq k_0 | i_k > i_{k-1}+1} (i_k - i_{k-1} - 1).$$

Clearly,

$$n\gamma_n(x_1, \pi) = k_0 \gamma_{k_0}(x_1, \pi') + \sum_{0 \leq k < k_0 | i_{k+1} > i_k + 1} (i_{k+1} - i_k - 1)\gamma_{i_k+1, i_{k+1}-1}(x_1, \pi).$$

Since Case 2 does not hold, $\gamma_{i_k+1, i_{k+1}-1}(x_1, \pi) < w - \varepsilon$ whenever $i_{k+1} > i_k + 1$. Since $n \geq k_0 \geq n_0$, $\gamma_n(x_1, \pi) \geq w - \varepsilon$. It follows that $\gamma_{k_0}(x_1, \pi') \geq w - \varepsilon$, as desired. □

COMMENT 3. The fact that the action set A is finite was used in the diagonal extraction argument in the proof of Proposition 4. However, the proof can be extended to compact metric action spaces provided the functions $a \mapsto r(\omega, a)$ and $a \mapsto q(\omega, a)$ are continuous in a, for each $\omega \in \Omega$.

To see why the diagonal extraction argument works in that case, take for every $n \in \mathbf{N}$ a finite subset $A_n \subset A$ such that for each $a \in A$ there is some $\bar{a}_n(a) \in A_n$ with

$$(7) \quad \sup_{\omega} |r(\omega, a) - r(\omega, \bar{a}_n(a))| < \varepsilon \quad \text{and} \quad \sup_{\omega} \|q(\omega, a) - q(\omega, \bar{a}_n(a))\| < \varepsilon/2^n.$$

Define for every $i \in \mathbf{N}$ the strategy π'_i by $\pi'_i(n) = \bar{a}_n(\pi_i(n))$. By (7), $|\gamma_n(x_1, \pi_i) - \gamma_n(x_1, \pi'_i)| < 2\varepsilon$. Since for each fixed n, $\{\pi'_i(n)\}_{i \in \mathbf{N}}$ is finite, one can apply the diagonal extraction argument to $\{\pi'_i\}_{i \in \mathbf{N}}$, and get a strategy π' such that every prefix of π' is a prefix of infinitely many π'_i's. Then π'_i is 3ε-optimal.

6. The general case. This section is devoted to the proof of Theorem 1. At first we follow the same path as in the proof of Theorem 2. However, since now the signal set is not degenerate, the posterior distribution at stage N_i depends on the signals the decision maker received. Hence, before the process starts, the decision maker who follows some strategy has a probability distribution over the possible posteriors he may have at stage N_i. We are thus forced to work with the space $\Delta(\Delta(\Omega))$, which is no longer finite dimensional. The proof will be amended to deal with this difficulty.

Fix $\varepsilon > 0$ once and for all. Denote $w = \limsup_{n\to\infty} v_n(x_1)$, and let (n_i) be an increasing subsequence such that $\lim_{i\to\infty} v_{n_i}(x_1) = w$ and $|w - v_{n_i}(x_1)| < \varepsilon$ for every $i \in \mathbf{N}$.

For each $i \in \mathbf{N}$, let σ_i be a pure optimal strategy in the n_i-stage problem (that satisfies the condition of Lemma 1), and let $N_i \leq 1 + (1-\varepsilon)n_i$ be the integer obtained by applying Proposition 2 to σ_i, ε and n_i.

Recall that y_{N_i} is the posterior distribution over Ω at stage N_i, given the history up to that stage. Since A and S are finite, y_{N_i} may take only finitely many values.

We denote by p_i the distribution of y_{N_i} when the strategy σ_i is followed: p_i has finite support $\operatorname{supp}(p_i)$ and

$$p_i[y] = \mathbf{P}_{x_1,\sigma_i}(y_{N_i} = y) \qquad \text{for each } y \in \Delta(\Omega).$$

COMMENT 4. A natural idea is to repeat the proof of the previous section, by using p_i instead of y_i, that is, by dealing with the auxiliary state space $\Delta(\Delta(\Omega))$. Observe that $\Delta(\Delta(\Omega))$ is no longer finite-dimensional but is compact in the w^*-topology, which is a metric topology. Let d be a corresponding metric. The proof of the previous section would go through if one was able to prove the following Lipschitz property:

$$\text{for every } p, p' \in \Delta(\Delta(\Omega)),\ \sigma \text{ and } n \in \mathbf{N}, \qquad |\gamma_n(p,\sigma) - \gamma_n(p',\sigma)| \leq d(p,p'),$$

where $\gamma_n(p,\sigma)$ denotes the expectation of $\gamma_n(x,\sigma)$ under p. However, it is not clear that this condition holds. We therefore choose a different route, which involves a discretization of $\Delta(\Omega)$, and uses the Lipschitz condition expressed in Lemma 1.

Let $\mathcal{T}$ be a fixed finite partition of $\Delta(\Omega)$ into sets of diameter smaller than ε. By Lemma 1, given $T \in \mathcal{T}$, $x, x' \in T$, a strategy σ and $n \in \mathbf{N}$, one has

$$|\gamma_n(x,\sigma) - \gamma_n(x',\sigma)| < \varepsilon. \tag{8}$$

Given $p \in \Delta(\Delta(\Omega))$ with finite support, we denote by $\hat{p}$ the probability induced by p on $\mathcal{T}$:

$$\hat{p}[T] = \sum_{x\in\operatorname{supp}(p)\cap T} p[x] \qquad \forall\, T \in \mathcal{T}.$$

Since $\mathcal{T}$ is a finite partition, there is a subsequence of $(\hat{p}_i)_{i\in\mathbf{N}}$ that converges to a limit $\hat{p}$. We still denote this subsequence by $(\hat{p}_i)_{i\in\mathbf{N}}$. We assume, moreover, that the support of $\hat{p}_i$ is independent of i, and that, for every $i \in \mathbf{N}$, $\|\ \hat{p}_i - \hat{p}\ \|_1 < \varepsilon/2$, where $\|x\|_1 = \sum_{k=1}^n |x_k|$ for $x \in \mathbf{R}^n$. In particular, $\|\hat{p}_i - \hat{p}_1\|_1 < \varepsilon$ for every $i \in \mathbf{N}$.

In the case of no signals, we defined a strategy π_i as: follow σ_1 up to stage N_1, then switch to the sequence of actions prescribed by σ_i after stage N_i. There is a small difficulty to proceed in a similar way here. The action that σ_i plays in stage N_i depends on y_{N_i}. However, the possible distributions at stage N_1 need not be the same as the possible distributions at stage N_i. Thus, one needs to define a map that associates to the *true* distribution y_{N_1} held at stage N_1 a *fictitious* value for y_{N_i}. The solution is simply to select a fictitious distribution x according to the conditional distribution $p_i[\cdot|T(y_{N_1})]$, where, given $y \in \Delta(\Omega)$, $T(y)$ is the element of $\mathcal{T}$ that contains y.

In other words, π_i follows σ_1 up to stage N_1. Denote by T the element of $\mathcal{T}$ that contains y_{N_1}. Choose now $y' \in T$ by $p_i[\cdot|T]$, the conditional distribution under σ_i at stage N_i. In particular, y' is a feasible posterior under σ_i at stage N_i, and so there is some history h of length N_i such that $y' = y_{N_i}(h)$. From stage N_1 and on, π_i follows $\sigma_i(h)$—it replaces the actual history up to stage N_1 by a fictitious history of length N_i.

To formalize this idea we need additional notation. For each $x \in \Delta(\Omega)$, we define the strategy $\sigma_i^{N_i}[x]$ induced by σ_i after stage N_i, given the distribution x, as follows. For each history $(a'_1, s'_1, \dots, a'_m, s'_m)$, we set

$$\sigma_i^{N_i}[x](a'_1, s'_1, \dots, a'_m, s'_m) = \sigma_i(a_1, s_1, \dots, a_{N_i-1}, s_{N_i-1}, a'_1, s'_1, \dots, a'_m, s'_m),$$

where $(a_1, s_1, \dots, a_{N_i-1}, s_{N_i-1})$ is any sequence in H_{N_i} such that

$$y_{N_i}(a_1, s_1, \dots, a_{N_i-1}, s_{N_i-1}) = x.$$

Since $\sigma_i(h_n)$ is a function of n and $y_n(h_n)$, this is independent of the particular sequence $(a_1, s_1, \dots, a_{N_i-1}, s_{N_i-1})$. (If no such sequence exists, the definition of $\sigma_i^{N_i}[x]$ is irrelevant.)

We now define, for every $i \in \mathbf{N}$, a strategy π_i as follows:

- Follow σ_1 up to stage $N_1 - 1$.
- If $p_i[T(y_{N_1})] = 0$, continue in an arbitrary way.
- Otherwise, choose y' according to $p_i[\cdot|T(y_{N_1})]$, and continue with $\sigma_i^{N_i}[y']$.

Observe that the definition of π_i involves choosing at stage N_1 a pure strategy at random, so that π_i is a mixed strategy. By Kuhn's theorem [8] we may view it as a behavioral strategy.

PROPOSITION 5. *For any m such that $(N_1 - 1)/\varepsilon < m \le N_1 + n_i - N_i$, one has*

$$\gamma_m(x_1, \pi_i) > w - 5\varepsilon.$$

PROOF. By the definition of π_i (by convention, $p_i[x \mid T] = 0$ as soon as $p_i[T] = 0$),

$$\gamma_m(x_1, \pi_i) \geq \frac{N_1 - 1}{m}\gamma_{N_1-1}(x_1, \sigma_1) + \frac{m - N_1 + 1}{m} \sum_{y \in \Delta(\Omega)} \sum_{x \in T(y)} p_1[y] p_i[x|T(y)] \gamma_{m-N_1+1}\big(y, \sigma_i^{N_i}[x]\big),$$

with equality if $p_i[T(y)] > 0$ for every $y \in \Delta(\Omega)$ such that $p_1[y] > 0$. Since payoffs are nonnegative, and by the assumption on m,

$$\gamma_m(x_1, \pi_i) \geq (1 - \varepsilon) \sum_{y \in \Delta(\Omega)} \sum_{x \in T(y)} p_1[y] p_i[x|T(y)] \gamma_{m-N_1+1}\big(y, \sigma_i^{N_i}[x]\big).$$

If $x, y \in \Delta(\Omega)$ belong to the same element of $\mathcal{T}$, one has by (8)

$$\big|\gamma_{m-N_1+1}\big(y, \sigma_i^{N_i}[x]\big) - \gamma_{m-N_1+1}\big(x, \sigma_i^{N_i}[x]\big)\big| \leq \varepsilon.$$

Therefore,

$$\gamma_m(x_1, \pi_i) \geq (1 - \varepsilon) \sum_{T \in \mathcal{T}} \hat{p}_1[T] \sum_{x \in T} p_i[x|T] \gamma_{m-N_1+1}\big(x, \sigma_i^{N_i}[x]\big) - \varepsilon. \tag{9}$$

Since $\|\hat{p}_i - \hat{p}_1\|_1 < \varepsilon$,

$$\begin{aligned} &\sum_{T \in \mathcal{T}} \hat{p}_1[T] \sum_{x \in T} p_i[x|T] \gamma_{m-N_1+1}\big(x, \sigma_i^{N_i}[x]\big) \\ &\quad \geq \sum_{x \in \Delta(\Omega)} p_i[x] \gamma_{m-N_1+1}\big(x, \sigma_i^{N_i}[x]\big) - \varepsilon \qquad (10) \\ &\quad = \gamma_{N_i, m-N_1+N_i}(x_1, \sigma_i) - \varepsilon. \end{aligned}$$

By (9), (10) and (5) we get

$$\begin{aligned} \gamma_m(x_1, \pi_i) &\geq (1 - \varepsilon)\gamma_{N_i, m-N_1+N_i}(x_1, \sigma_i) - 2\varepsilon \\ &\geq (1 - \varepsilon)\big(v_{n_i}(x_1) - \varepsilon\big) - 2\varepsilon \\ &> (1 - \varepsilon)(w - 2\varepsilon) - 2\varepsilon > w - 5\varepsilon. \end{aligned}$$

□

The last step is to construct from the sequence $(\pi_i)_{i \in \mathbf{N}}$, using a diagonal extraction argument, a strategy π that is 5ε-optimal. In this step we use the representation of π_i as a behavioral strategy. Let $n \geq 1$ be given. Since H_n is finite, there exists a sequence $(i_n(j))_{j \in \mathbf{N}}$ such that $\lim_{j \to \infty} \pi_{i_n(j)}(h)$ exists for every $h \in H_n$. We denote by $\pi(h)$ the limit. Without loss of generality, we may assume that $(i_{n+1}(j))_j$ is a subsequence of $(i_n(j))_j$ for each n. Clearly, for each $n \in \mathbf{N}$,

$$\gamma_n(x_1, \pi) = \lim_{j \to \infty} \gamma_n(x_1, \pi_{i_n(j)}).$$

By Proposition 5, $\gamma_n(x_1, \pi) > w - 5\varepsilon$, for every $n > (N_1 - 1)/\varepsilon$. Hence Theorem 1 is proved. □

We conclude by discussing several extensions.

COMMENT 5. The extension to a compact set of actions also holds in the general case, under the same conditions as in the case of no signals, as discussed above.

COMMENT 6. The extension to MDP with finite Ω, A and a countable set of signals S is straightforward. Indeed, given $\varepsilon > 0$, there exist finite subsets S_n^* of S such that, given any strategy σ and any initial distribution $x_1 \in \Delta(\Omega)$,

$$\mathbf{P}_{x_1,\sigma}(s_n \notin S_n^* \text{ for some } n) \leq \varepsilon/2^n.$$

The proof then essentially reduces to the case of a finite set of signals.

COMMENT 7. The extension to MDP with finite A and countable Ω does not hold, even when S is a singleton. Indeed, there are examples, see [10] for instance, of an MDP with finite A, countable Ω and deterministic transitions, that have no value. For such MDP, the sequence of past actions enables the decision maker to recover the current state of the MDP. Hence the assumption of partial observation is irrelevant.

COMMENT 8. Our proof works in the case of MDPs with a compact metric space Ω, and finite action set A and signal set S, as long as (8) holds.

Acknowledgments. We thank an Associate Editor and two anonymous referees. Their comments substantially improved the presentation.

REFERENCES

[1] ALTMAN, E. (2001). Applications of Markov decision processes in communication networks: a survey. In *Handbook of Markov Decision Processes: Methods and Applications* (E. Feinberg and A. Shwartz, eds.). Kluwer, Boston.

[2] ARAPOSTATHIS, A., BORKAR, V. S., FERNÁNDEZ-GAUCHERAND, E., GHOSH, M. K. and MARCUS, S. I. (1993). Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control Optim.* **31** 282–344.

[3] BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.

[4] BORKAR, V. S. (1988). Control of Markov chains with long-run average cost criterion. In *Stochastic Differential Systems, Stochastic Control Theory and Applications* (W. Fleming and P. L. Lions, eds.) 57–77. Springer, Berlin.

[5] BORKAR, V. S. (1991). *Topics in Controlled Markov Chains.* Longman, Essex.

[6] FERNÁNDEZ-GAUCHERAND, E., ARAPOSTATHIS, A. and MARCUS, S. I. (1989). On partially observable Markov decision processes with an average cost criterion. In *Proceedings of the 28th IEEE Conference on Decision and Control* 1267–1272. IEEE Press, New York.

[7] KALLENBERG, L. (2001). Finite state and action MDPs. In *Handbook of Markov Decision Processes: Methods and Applications* (E. Feinberg and A. Shwartz, eds.) 21–30. Kluwer, Boston.

[8] KUHN, H. W. (1953). Extensive games and the problem of information. In *Contributions to the Theory of Games II* (H. W. Kuhn and A. W. Tucker, eds.) 193–216. Princeton Univ. Press.

[9] LANE, D. E. (1989). A partially observable model of decision making by fishermen. *Oper. Res.* **37** 240–254.

[10] LEHRER, E. and SORIN, S. (1992). A uniform Tauberian theorem in dynamic programming. *Math. Oper. Res.* **17** 303–307.

[11] MITRA, T., RAY, D. and ROY, R. (1991). The economics of orchards: an exercise in point-input, flow-output capital theory. *J. Econom. Theory* **53** 12–50.

[12] MONAHAN, G. E. (1982). A survey of partially observable Markov decision processes: theory, models, and algorithms. *Management Sci.* **28** 1–16.

[13] PUTERMAN, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, New York.

[14] RHENIUS, D. (1974). Incomplete information in Markovian decision models. *Ann. Statist.* **2** 1327–1334.

[15] SAWARAGI, Y. and YOSHIKAWA, T. (1970). Discrete time Markovian decision processes with incomplete state observation. *Ann. Math. Statist.* **41** 78–86.

[16] SENNOTT, L. I. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems*. Wiley, New York.

[17] YUSHKEVICH, A. A. (1976). Reduction of a controlled Markov model with incomplete date to a problem with complete information in the case of Borel state and control spaces. *Theory Probab. Appl.* **21** 153–158.

D. ROSENBERG
LABORATOIRE D'ANALYSE GEOMETRIE
ET APPLICATIONS
INSTITUT GALILÉE
UNIVERSITÉ PARIS NORD
AVENUE JEAN BAPTISTE CLÉMENT
93430 VILLETANEUSE
FRANCE
E-MAIL: dinah@math.univ-paris13.fr

E. SOLAN
KELLOGG SCHOOL OF MANAGEMENT
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS 60208
AND
SCHOOL OF MATHEMATICAL SCIENCES
TEL AVIV UNIVERSITY
TEL AVIV 69978
ISRAEL
E-MAIL: eilons@post.tau.ac.il

N. VIEILLE
ECOLE POLYTECHNIQUE
AND
DÉPARTEMENT FINANCE ET ECONOMIE
HEC
1, RUE DE LA LIBÉRATION
78 351 JOUY-EN-JOSAS
FRANCE
E-MAIL: vieille@hec.fr