Blackwell Optimality in Markov Decision Processes with Partial Observation

Dinah Rosenberg *

and

Eilon Solan †

and

Nicolas Vieille ‡

April 6, 2000

Abstract

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

*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

†Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston IL 60208. e-mail: e-solan@nwu.edu

‡GRAPE, Université Montesquieu-Bordeaux 4, and Laboratoire d'Econométrie de l'Ecole Polytechnique, 1 rue Descartes, 75 005 Paris, France. e-mail: vieille@poly.polytechnique.fr

1 Introduction

A well-known result by Blackwell [1] states that, in any Markov Decision Process (MDP thereafter) 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 address the problem of existence of Blackwell optimal strategies for finite MDP with partial observation; that is, for finite MDP's 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. We prove that, in any such MDP, 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.

The standard approach to MDP's with partial observation 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 [7], [8] and [9]. 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 the underlying state in Ω). A commonly used criterion is the long-run average cost criterion, see, e.g., [2], [3]. It is well-known that optimal strategies for this criterion do not exist in general MDP's with Borel state space. Hence one imposes assumptions which guarantee the existence of optimal strategies. These assumptions usually have the flavor of an irreducibility condition that one imposes on the transition function of the MDP. For MDP's that arise from a MDP with partial observation, these conditions may be difficult to interpret in terms of the underlying data; see for instance Assumption 7.2, p. 329 in [6].

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 $r : \Omega \times A \to \mathbf{R}$, (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. W.l.o.g., we assume that $0 \leq r(\omega, a) \leq 1$ for every $(\omega, a) \in \Omega \times A$.

An initial state ω_1 is drawn according to x_1. At every stage n the decision maker chooses an action $a \in A$, 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}.

Thus, the information available to the decision maker at stage n is the finite sequence $a_1, s_1, a_2, s_2, \ldots, a_{n-1}, s_{n-1}$ and a *behavior 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 $H = \cup_{n \geq 1} H_n$ 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 over H_∞ induced by H_n.

Each strategy σ and every initial distribution x_1 induce 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.

We let

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

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 process. We simply write v_n where there is no risk of confusion about the initial distribution.

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

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

and the discounted value by

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

Definition 1 *$v \in \mathbf{R}$ is the (uniform) value of the MDP with p.o. (with initial probability distribution x_1) if $v = \lim_{n\to\infty} v_n = \lim_{\lambda\to 1} v_\lambda$ and, for every $\varepsilon > 0$, there exists a strategy σ, a positive integer $N_0 \in \mathbf{N}$, and $\lambda_0 \in (0,1)$ such that:*

$$\gamma_n(x_1,\sigma) \geq v_n - \varepsilon,\ \forall n \geq N_0 \tag{1}$$
$$\gamma_\lambda(x_1,\sigma) \geq v_\lambda - \varepsilon,\ \forall \lambda \geq \lambda_0 \tag{2}$$

Our first main result is that the value always exists.

Theorem 2 *If Ω, A and S are finite, then v exists.*

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 law 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 *law* 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 the belief $y_n(h_n)$ held at stage n.

If $|S| = 1$, the ε-optimal strategies can be chosen to be pure and stationary.

Theorem 3 *If Ω and A are finite, and $|S| = 1$, then for every $\varepsilon > 0$ there exists a pure stationary ε-optimal strategy.*

Comment: It might seem that stationarity is an extremely desirable requirement. However, it may well be the case that the decision maker cannot hold twice the same belief over time. In such a case, the stationarity requirement is empty.

Comment: It is not clear that the existence of a *pure* ε-optimal strategy follows from the existence of ε-optimal strategies (*i.e.*, from Theorem 2). The

reason is the following. By Kuhn's theorem [4], given x_1 and a strategy σ, there exists a mixed strategy π, i.e., a probability distribution over pure strategies, such that the probability distribution over H_∞ obtained by first choosing a pure strategy f according to π, and then following f, coincides with $\mathbf{P}_{x_1,\sigma}$. In particular, given $n \geq 1$, 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.

3 An example

Define a MDP with no signals as follows. Set $\Omega = \{\underline{\omega}, \overline{\omega}\}$, and $A = \{\underline{a}, \overline{a}\}$. The transition rule q is given by

$$q(\bar{\omega}|\bar{\omega}, a) = 1 \text{ for each } a$$
$$q(\underline{\omega}|\underline{\omega}, \bar{a}) = 1, q(\underline{\omega}|\underline{\omega}, \underline{a}) = \frac{1}{2}.$$

The payoff function r is given by

$$r(\bar{\omega}, \bar{a}) = 1, \text{ and } r(\omega, a) = 0 \text{ otherwise.}$$

The MDP starts from state $\underline{\omega}$. We identify a probability distribution over Ω with the probability assigned to $\underline{\omega}$.

Observe that the state $\bar{\omega}$ is absorbing. Observe also that: whenever the player chooses $\bar{a}$, the current state does not change, hence the belief remains the same; whenever the player choose $\underline{a}$, the current belief (i.e., the probability of being in $\underline{\omega}$) is divided by two.

The uniform value of this MDP is equal to one. Indeed, given $\varepsilon > 0$, let σ be the (stationary) strategy that plays $\underline{a}$ in the first $N = \lfloor \log_2 \varepsilon \rfloor + 2$ stages, and plays $\bar{a}$ 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, $\liminf_{n\to\infty} \gamma_n(\sigma) = \liminf_{\lambda\to 1} \gamma_\lambda(\sigma) > 1 - \varepsilon$. Since $v_n \leq 1$, and $v_\lambda \leq 1$, the uniform value is indeed equal to 1. This implies $\lim_{\lambda\to 1} v_\lambda = \lim_{n\to\infty} v_n = 1$.

We now claim that there is no Blackwell optimal strategy. By Kuhn's theorem, it is enough to prove that there is no pure Blackwell optimal strategy. Let $\sigma = (a_n)_{n\in\mathbf{N}}$ be a pure strategy. We distinguish three (non-exclusive) cases.

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

In that case, the sequence (y_n) is constant from stage N on. Therefore, $\lim_{n\to\infty}\gamma_n(\sigma) = \lim_{\lambda\to 1}\gamma_\lambda(\sigma) = 1 - y_N < 1$. In particular, $\gamma_\lambda(\sigma) < v_\lambda$ for λ close to one.

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

In that case, $\mathbf{E}_\sigma\left[r(\omega_n, a_n)\right] = 0$ for each $n \geq N$. Therefore, $\lim_{n\to\infty}\gamma_n(\sigma) = \lim_{\lambda\to 1}\gamma_\lambda(\sigma) = 0$.

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

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

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

A natural question arises. Does there exist a strategy that is Blackwell ε-optimal for each $\varepsilon > 0$? We claim that there is such a pure strategy, but no stationary one. Indeed, let $\sigma = (a_n)_{n\in\mathbf{N}}$ be a pure stationary strategy. Since $y_{n+1} = y_n$ whenever $a_n = \bar{a}$, the stationarity of σ implies that $a_{n+1} = \bar{a}$ as soon as $a_n = \bar{a}$. This implies that the sequence (a_n) is eventually constant, i.e., 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 $\underline{A} = \{n \in \mathbf{N}, a_n = \underline{a}\}$ of $\mathbf{N}$ is infinite and has density zero. Since $\underline{A}$ is infinite, the sequence (y_n) converges to zero under σ. Therefore,

$$\lim_{n\to\infty, n\notin\underline{A}} \mathbf{E}_\sigma\left[r(\omega_n, a_n)\right] = 1. \tag{3}$$

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

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 in stage n depends only on n and y_n.

Lemma 4 *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 $y_n(h_n)$.*

Proof. Let a strategy σ be given, and define a strategy $\hat{\sigma}$ as follows. In stage $n \geq 1$, it plays $a \in A$ with the probability $\mathbf{P}_{x,\sigma}(a_n = a|y_1, ..., y_n)$. Since y_n is a sufficient statistic about ω_n, it is easy to check that $\gamma_N(x_1, \sigma) = \gamma_N(x_1, \hat{\sigma})$. Observe that $\hat{\sigma}(h_n)$ depends only on n and $y_n(h_n)$. Using Kuhn's Theorem, there exists a pure strategy σ_N such that $\gamma_N(x_1, \sigma_N) \geq \gamma_N(x_1, \hat{\sigma})$. The result follows. ■

Whenever in the sequel we refer to optimal strategies in the n-stage process, we mean a pure strategy that satisfies the two conditions in Lemma 4.

Given $m < n$, we denote by

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

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

Proposition 5 *Let $x, x' \in \Delta(\Omega)$. For every strategy σ and every $m < n$,*

$$|\gamma_{m,n}(x, \sigma) - \gamma_{m,n}(x', \sigma)| \leq \|x - x'\|.$$

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

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

In particular, $\mathbf{E}_{x,\sigma}[r(s_n, a_n)] = \sum_{\omega \in \Omega} x(\omega)\mathbf{E}_{\omega,\sigma}[r(s_n, a_n)]$. The result follows. ■

For simplicity, we write $\gamma_n(\sigma)$ and $\gamma_{m,n}(\sigma)$ instead of $\gamma_n(x_1, \sigma)$ and $\gamma_{m,n}(x_1, \sigma)$ whenever there is no possible confusion about x_1.

Comment: We claim here that to prove that v is the value, it is enough to prove that $v = \lim_{n\to\infty} v_n$ and (1) holds. Since Ω is finite, Proposition 5 implies that (v_n) converges to v uniformly over $\Delta(\Omega)$. Hence, by Lehrer and Sorin [5], (v_λ) converges uniformly to v. Moreover, one can show 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.

Proposition 6 *Let σ, $\varepsilon > 0$ and $n \in \mathbf{N}$ be given, and set*

$$N = \inf \{k \in \mathbf{N}, \text{ such that } \gamma_m(\sigma) \geq \gamma_n(\sigma) - \varepsilon \text{ for every } k \leq m \leq n\}. \quad (4)$$

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

$$\gamma_{N,m}(\sigma) \geq \gamma_n(\sigma) - \varepsilon \text{ for every } N \leq m \leq n. \quad (5)$$

Given $\varepsilon > 0$ and σ, let N_n denote the integer associated with n in (4). Observe that $\lim_{n\to\infty}(n - N_n) = +\infty$. This Proposition has the same flavor as Proposition 2 in [5].

Proof. Clearly, $N \leq n$. Note that if $N > 1$ then $\gamma_{N-1}(\sigma) < \gamma_n(\sigma) - \varepsilon$.

We first show that $N \leq 1 + (1-\varepsilon)n$. Indeed, otherwise, $N > 1$, hence $\gamma_{N-1}(\sigma) < \gamma_n(\sigma) - \varepsilon$. Since payoffs are bounded by 1,

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

a contradiction.

Next we show that (5) holds. Fix an integer m such that $N \leq m \leq n$. If $N = 1$, one has $\gamma_{N,m}(\sigma) = \gamma_m(\sigma) \geq \gamma_n(\sigma) - \varepsilon$. If $N > 1$, $\gamma_{N-1}(\sigma) < \gamma_n(\sigma) - \varepsilon$, while $\gamma_m(\sigma) \geq \gamma_n(\sigma) - \varepsilon$. It follows that $\gamma_{N,m}(\sigma) \geq \gamma_n(\sigma) - \varepsilon$. ■

5 The Case of 'No Signals'

This section is devoted to the proof of Theorem 3. Thus, we assume that no signal is available. The initial distribution x_1 is fixed throughout the section.

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 ${}^m\sigma = (a_1, ..., a_m)$.

Define $w = \limsup_{n\to\infty} v_n$, and fix $\varepsilon > 0$. Let $(n_i)_{i\in\mathbf{N}}$ be a subsequence such that $\lim_{i\to\infty} v_{n_i} = w$, and $|v_{n_i} - w| < \varepsilon/2$ for every $i \in \mathbf{N}$. Let σ_i be a pure optimal strategy in the n_i-stage problem (that satisfies the two conditions of Lemma 4.) Thus, $\gamma_{n_i}(\sigma_i) = v_{n_i}$.

Given $i \in \mathbf{N}$, we let $N_i \leq 1 + (1-\varepsilon)n_i$ be the integer obtained by applying Proposition 6 to n_i. Possibly by taking a subsequence, we may assume w.l.o.g. that $N_1 \leq N_i$ for each i.

We let $y_i = y_{N_i}(\sigma_i)$ be the posterior distribution over states induced by σ_i at stage N_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

$$\|y_i - y\| < \varepsilon, \text{ for each } i \in \mathbf{N}.$$

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,

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

Set $m_i = N_1 + n_i - N_i$. Since $N_1 \leq N_i$, one has $m_i \leq n_i$. Note that $\liminf_{i\to\infty} m_i = +\infty$.

Proposition 7 *If m satisfies $(N_1 - 1)/\varepsilon < m \leq m_i$ then*

$$\gamma_m(\pi_i) \geq w - 4\varepsilon.$$

Proposition 7 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. Fix an integer m such that $(N_1 - 1)/\varepsilon < m \leq m_i$. 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, $(m - N_1 + 1)/m \geq 1 - \varepsilon$. Since $\|y_1 - y_i\| < \varepsilon$, we get by Proposition 5, and since $\gamma_{N_1-1}(\pi_i) \geq 0$,

$$\gamma_m(x_1, \pi_i) \geq (1-\varepsilon)\left(\gamma_{N_i,m-N_1+N_i}(x_1, \pi_i) - \varepsilon\right).$$

Since $N_1 \leq N_i$, $m - N_1 + N_i \leq n_i$, hence $\gamma_{N_i,m-N_1+N_i}(y_i, \pi_i) \geq w - 2\varepsilon$. The result follows. ∎

Proposition 8 *In the case $|S| = 1$, the uniform value exists.*

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 $m > (N_1 - 1)/\varepsilon$,

$$\gamma_m(\pi) \geq w - 4\varepsilon.$$

In particular, $v_m \geq w - 4\varepsilon$. Since $\varepsilon > 0$ is arbitrary, one has $w = \lim_{n\to\infty} v_n$ and π is a 4ε-optimal strategy. ■

Proof of Theorem 3. Let $\pi = (a_1, a_2, \dots)$ be a pure ε-optimal strategy; that is, for some $n_0 \in \mathbf{N}$, $\gamma_n(\pi) \geq w - \varepsilon$ for every $n \geq n_0$. Let $y_n = y_n(\pi)$ be the posterior distribution at stage n.

Case 1: $(y_n)_{n\in\mathbf{N}}$ is eventually periodic; that is, there exists $n_1 \in \mathbf{N}$ and $d \in \mathbf{N}$ such that $y_n = y_n + d$ for every $n \geq n_1$.

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

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

It follows that there exist $n_2 \leq n_3$ such that (i) $y_{n_2} = y_{n_3}$, (ii) $y_i \neq y_j$ for every $n_2 \leq i < j < n_3$, and (iii) $\gamma_{n_2,n_3-1}(\pi) \geq w - \varepsilon$.

Let $Y = \{y_n, n = 1, \dots, n_3\}$ be the set of all posterior distributions in the first n_3 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 $n \in \{1, \dots, n_3 - 1\}$. Thus we connect with an edge any two consecutive elements in the finite sequence $(y_n)_{n=1}^{n_3}$.

Clearly there is a path from y_1 to any $y \in Y$. Let $y_1 = y_{i_1}, y_{i_2}, \dots, y_{i_k}$ be a shortest path that connects y_1 to the set $\{y_{n_2}, y_{n_2}+1, \dots, y_{n_3}\}$. 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

$$\pi' = (a_{i_1}, a_{i_2}, \dots, a_{i_k-1}, a_{n_2}, a_{n_2+1}, \dots, a_{n_3-1}, a_{n_2}, a_{n_2+1}, \dots, a_{n_3-1}, \dots).$$

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 $(y_{n_2}(\pi), ...y_{n_3-1}(\pi), y_{n_2}(\pi), ..., y_{n_3-1}(\pi), ...)$. Each of the posteriors $y_n(\pi')$, $n < k$ appears only once, hence π' is stationary. Since $\gamma_{n_2,n_3-1}(\pi) \geq w - \varepsilon$, we have $\gamma_n(\pi') \geq w - 2\varepsilon$ for every $n \geq k(n_3 - n_2)/\varepsilon$.

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

Define the strategy $\pi' = (a_1, a_2, \dots, a_{n_1}, a_{n_1+1}, \dots, a_{n_2-1}, a_{n_1}, \dots, a_{n_2-1}, \dots)$. 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, $\gamma_n(\pi) \geq w - \varepsilon$, it follows that there exist $n_1 < n_2$ such that $y_{n_1} = y_{n_2} = y$ and $\gamma_{n_1,n_2-1}(\pi) \geq w - \varepsilon$. Apply now Case 2.

Case 4: None of the above hold.

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 belief, i_2 is the last occurrence of the belief held in 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 $\pi' = (a_{i_1}, a_{i_2}, a_{i_3}, \dots)$. Since $y_{i_{k+1}} = y_{i_k+1}$, it follows by induction that

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

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

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

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

Clearly,

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

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

Comment. The fact that the action set A is finite was used in the diagonal extraction argument in the proof of Proposition 8. 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 $\bar{a}(a) \in A_n$ with

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

Define for every $i \in \mathbf{N}$ the strategy π'_i by $\pi'_i(n) = \bar{a}(\pi_i(n))$. By (7), $|\gamma_n(\pi_i) - \gamma_n(\pi'_i)| < 2\varepsilon$. 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 2. At first we follow the same path as in the proof of Theorem 3. 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 $w = \limsup_{n \to \infty} v_n$, and let (n_i) be a subsequence such that $\lim_{i \to \infty} v_{n_i} = w$ and $|w - v_{n_i}| < \varepsilon$ for every $i \in \mathbf{N}$.

For each $i \in \mathbf{N}$, let σ_i be an optimal strategy in the n_i-stage MDP (that satisfies the two conditions of Lemma 4), and let $N_i \leq 1 + (1 - \varepsilon)n_i$ be the integer obtained by applying Proposition 6 to n_i.

We assume w.l.o.g. that $N_1 \leq N_i$ for each 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 law of y_{N_i} when the strategy σ_i is followed (under $\mathbf{P}_{\sigma_i}$): p_i has finite support $\mathrm{supp}(p_i)$ and

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

Comment. A natural idea is to repeat the proof of the previous section, by using the law of the belief as relevant state variable, i.e. 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:

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

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 5.

Let $\mathcal{T}$ be a finite partition of $\Delta(\Omega)$ into sets of diameter smaller than ε. By Lemma 5, given $T \in \mathcal{T}$, $x, x' \in T$, a strategy σ and every $n \in \mathbf{N}$, 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 $\mathcal{T}$:

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

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 for every $i \in \mathbf{N}$, $\| \hat{p}_i - \hat{p} \| < \varepsilon/2$. In particular, $\|\hat{p}_i - \hat{p}_1\| < \varepsilon$ 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. We will proceed in a similar way here. There is however a small difficulty. The action that σ_i plays in stage N_i depends on the belief y_{N_i}. Therefore, one needs to define a map that associates to the *true* belief y_{N_1} held at stage N_1 a *fictitious* value for y_{N_i}. Indeed, the possible beliefs in stage N_1 need not be the same as the possible beliefs in stage N_i. The solution is simply to select a fictitious belief x according to the conditional distribution $p_i\left[\cdot | T(y_{N_1})\right]$, where, given $y \in \Delta(\Omega)$, $T(y)$ is the element of $\mathcal{T}$ that contains y.

We need an 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 belief 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 $(a_1, s_1, \ldots, a_{N_i-1}, s_{N_i-1})$ is any sequence in H_{N_i} such that

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

Since σ_i is stationary, this is independent of the particular sequence $(a_1, s_1, \ldots, a_{N_i-1}, s_{N_i-1})$. (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 x according to $p_i[\cdot \mid T(y_{N_1})]$, and continue with $\sigma_i^{N_i}[x]$.

Observe that the definition of π_i involves choosing at stage N_1 a pure strategy at random. Such a strategy is called a mixed strategy. By Kuhn's theorem [4], there is a behavior strategy that induces the same probability distribution over H_∞ as π_i. We may therefore view π_i as a behavior strategy.

Proposition 9 *For any m such that $N_1/\varepsilon \leq m \leq N_1 + n_i - N_i$, one has*

$$\gamma_m(\pi_i) \geq w - 5\varepsilon.$$

Proof. By the definition of π_i, and since payoffs are bounded by 1:

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

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

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

Therefore

$$\gamma_m(\pi_i) \geq \frac{N_1 - 1}{m}\gamma_{N_1-1}(\sigma_1) \tag{9}$$

$$+ \frac{m - N_1 + 1}{m} \sum_{T \in \mathcal{T}} \hat{p}_1(T) \sum_{x \in T} p_i(x|T) \gamma_{m-N_1+1}(x, \sigma_i^{N_i}[x]) - \varepsilon. \tag{10}$$

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

$$\sum_{T\in\mathcal{T}} \hat{p}_1(T) \sum_{x\in T} p_i(x|T)\gamma_{m-N_1+1}(x, \sigma_i^{N_i}[x]) \geq \sum_{x\in\Delta(\Omega)} p_i(x)\gamma_{m-N_1+1}(x, \sigma_i^{N_i}[x]) - \varepsilon$$
$$\geq \gamma_{N_i, m-N_1+N_i}(\sigma_i) - \varepsilon$$

Since $m \geq N_1/\varepsilon$, substituting into (9) yields

$$\gamma_m(\pi_i) \geq (1-\varepsilon)\gamma_{N_i,m-N_1+N_i}(\sigma_i) - 2\varepsilon \geq w - 5\varepsilon.$$

■

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 6ε-optimal. 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. W.l.o.g., we may assume that $(i_{n+1}(j))_j$ is a subsequence of $(i_n(j))_j$ for each n. Clearly, for each n,

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

By Proposition 9, $\gamma_n(\pi) \geq w - 5\varepsilon$, for every $n \geq N_1/\varepsilon$. Hence Theorem 2 is proved.

We conclude by discussing several extensions.

Comment. 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. The extension to MDP with finite Ω, A and 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 $x_1 \in \Delta(\Omega)$,

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

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

Comment. The extension to MDP with finite A, countable Ω does not hold, even when S is a singleton. Indeed, there are examples, see [5] for instance, of 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. 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.

References

[1] D. Blackwell. Discrete dynamic programming. *Annals of Mathematical Statistics*, 33:719–726, 1962.

[2] V.S. Borkar. Control of markov chains with long-run average cost criterion. In W. Fleming and P.L. Lions, editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications, The IMA Volumes in Mathematics and Its Applications, Vol. 10*, pages 57–77. Springer-Verlag, Berlin, 1988.

[3] E. Fernandez-Gaucherand, A. Araposthatis, and S.I. Marcus. On partially observable markov decision processes with an average cost criterion. In *Proceedings of the 28th IEEE Conference on Decision and Control*, pages 1267–1272, Tampa, FL., 1989.

[4] H.W. Kuhn. Extensive games and the problem of information. In H.W. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games*. Annals of Mathematics Study 28, Princeton University Press, 1953.

[5] E. Lehrer and S. Sorin. A uniform tauberian theorem in dynamic prgramming. *Mathematics of Operations Research*, 17:303–307, 1992.

[6] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.

[7] D. Rhenius. Incomplete information in markovian decision models. *Annals of Statistics*, 2:1327–1334, 1974.

[8] S. Sawaragi and T. Yoshikawa. Discrete time markovian decision processes with incomplete state observation. *Annals of Mathematical Statistics*, 41:78–86, 1970.

[9] A.A. Yushkevich. Reduction of a controlled markov model with incomplete date to a problem with complete information in the cas of borel state and control spaces. *Theory of Probability and its Applications*, 21:153–158, 1976.