Sensitive optimality in stationary Markovian decision problems on a general state space

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 76-21

Sensitive optimality in stationary Markovian decision problems on a general state space

by

J. Wijngaard

Eindhoven, November 1976

The Netherlands

Sensitive optimality in stationary Markovian decision problems on a general state space.

J. Wijngaard

INTRODUCTION

In considering Markovian decision problems with no discounting the first interest is in general in the average costs. But if there are more average optimal strategies one can distinguish between these by considering the bias, the limit of the difference of the n-period costs and n times the average costs. An average optimal strategy which, among all average optimal strategies, minimizes the bias, is called sensitive optimal. Sensitive optimality is equivalent with 1-optimality (Blackwell [2]). Sensitive optimality and extensions are considered by Veinott [10], [11], Miller and Veinott [8] for a finite state space and by Hordijk and Sladky [7] for a countable state space.
In this paper we consider the existence of sensitive optimal strategies for problems on a general state space. Compactness of the space of strategies and continuity of the transition probability and the one-period costs on the space of strategies are used to derive sufficient conditions for the existence of sensitive optimal strategies.

1. Preliminaries

Let (V,Σ) be a measurable space. The linear space $B(V,\Sigma)$ is defined as the space of all complex valued bounded measurable functions on V. Let $||f|| := \sup_{u\in V} |f(u)|$ for all $f\in B(V,\Sigma)$, then $||.||$ is a norm on $B(V,\Sigma)$ and with this norm $B(V,\Sigma)$ in a Banach space.

A Markov process on (V,Σ) with transition probability P defines a bounded linear operator in $B(V,\Sigma)$ by

$$(Pf)(u) = \int_V f(v)P(u, dv), \quad f\in B(V,\Sigma)$$

The norm of this operator in $B(V,\Sigma)$ is denoted by $||P||$ and its spectrum by $\sigma(P)$. Since P is a Markov process, $1\in\sigma(P)$ and $\sigma(P)$ contains no points outside the unit circle

For $A \in \Sigma$ the sub-Markov process P_A is defined by

$$P_A(u, E) := P(u, A \cap E) \quad , u \in V, F \in \Sigma$$

Let $A \in \Sigma$, $B = V \backslash A$ and let Q be the embedded sub-Markov process of P on A, then

$$Q(u, E) = \sum_{n=0}^{\infty} (P_B^n P_A 1_E)(u) \quad , u \in V, E \in \Sigma$$

If $\lim_{n \to \infty} (P_B^n 1_V)(u) = 0$ for all $u \in V$ then Q is a Markov process.

Let c be a nonnegative measurable function. The pair (P, c) is called a <u>Markov process with costs</u>. If P is <u>quasi-compact</u> (satisfies the <u>Doeblin condition</u>) and c is bounded, the <u>average costs</u> $g := \lim_{n\to\infty} \frac{1}{n} \sum_{\ell=0}^{n-1} P^\ell c$ exist and the functions $w_m := \lim_{k\to\infty} \sum_{\ell=0}^{kd+m} P^\ell (c-g)$ exist for all m = 0, 1, 2, ... and for d equal to the period of P.

Let $v := \frac{1}{d} \sum_{m=0}^{d-1} w_m$, then v is a solution of $y = c-g + Py$ and if P has only one ergodic set this solution is unique upto a constant. The function v is called the <u>bias</u> of (P, c).

A <u>stationary Markovian decision problem</u> (SMD) is a set of Markov processes with costs $\{(P_\alpha, c_\alpha)\}$, $\alpha \in A$. The elements $\alpha \in A$ are called <u>strategies</u>. It is clear that if in a Markovian decision process only stationary policies are allowed, it can be interpreted as an SMD. An important property of an SMD is the <u>product property</u>.

An SMD satisfies the product property if for each $\alpha_1, \alpha_2 \in A$ and for each $F \in \Sigma$ there exists an $\alpha \in A$ such that

$$P_\alpha(u, E) = P_{\alpha_1}(u, E) \text{ and } c_\alpha(u) = c_{\alpha_1}(u) \quad \text{for } u \in F$$

$$P_\alpha(u, E) = P_{\alpha_2}(u, E) \text{ and } c_\alpha(u) = c_{\alpha_2}(u) \quad \text{for } u \in V \backslash F$$

This product property is always satisfied in Markovian decision processes, the actions in the different states may be chosen independently of each other.

If the product property holds it is possible to prove that for two arbitrary strategies, $\alpha_1, \alpha_2 \in A$ there exists a third strategy $\alpha \in A$ which is better than both. This is worked out in the next lemma.

Lemma 1. Let $\{(P_\alpha, c_\alpha)\}$, $\alpha \in A$ be an SMD with P_α quasi-compact and c_α bounded on V, uniform in α. Assume that the product property is satisfied. Let $\alpha_1, \alpha_2 \in A$ and $g_{\alpha_1}, g_{\alpha_2}$ and $v_{\alpha_1}, v_{\alpha_2}$ the corresponding average costs and bias. Then

i there exists a strategy $\alpha_0 \in A$ such that

$$g_{\alpha_0}(u) \le \min\{g_{\alpha_1}(u), g_{\alpha_2}(u)\} \quad \text{for all } u \in V$$

ii if α_1, α_2 are both average optimal then there exists a strategy $\alpha_0 \in A$ such that

$$v_{\alpha_0}(u) \le \min\{v_{\alpha_1}(u), v_{\alpha_2}(u)\} \quad \text{for all } u \in V$$

Proof. For the proof of the first part we refer to [12], section 4.1.3. Now let α_1, α_2 be two average optimal strategies, $g_{\alpha_1} = g_{\alpha_2} = g$.
Let $F := \{u \mid v_{\alpha_1}(u) < v_{\alpha_2}(u)\}$ and $G := V \setminus F$.
Let Q_{α_2} be the embedded sub-Markov process of P_{α_2} on F and Q_{α_1} the embedded sub-Markov process of P_{α_1} on G.
The strategy α_0 is chosen such that

$$P_{\alpha_0}(u, E) = P_{\alpha_1}(u, E),\ c_{\alpha_0}(u) = c_{\alpha_1}(u) \quad \text{for } u \in F$$

$$P_{\alpha_0}(u, E) = P_{\alpha_2}(u, E),\ c_{\alpha_0}(u) = c_{\alpha_2}(u) \quad \text{for } u \in G$$

The product property implies that there is such a strategy α_0 in A. Let R_{α_0} be the entry process of P_{α_0} on F, that means that R_{α_0} is the sub-Markov process which describes the state of the system each time the set F is entered,

$$R_{\alpha_0}(u, E) = Q_{\alpha_2}(u, E) \qquad , u \in G$$

$$R_{\alpha_0}(u, E) = (Q_{\alpha_1} Q_{\alpha_2})(u, E),\ u \in F$$

Define $v_{\alpha_1 n \alpha_2}$ as the bias of the (non-stationary) strategy which applies α_0 until the set F is entered for the n^{th} time and from then on the strategy α_1.

Consider first the case that α_0 has only one invariant probability π_{α_0}.

If $\pi_{\alpha_0}(F) > 0$ and $\pi_{\alpha_0}(G) > 0$ then Q_{α_2} and Q_{α_1} are Markov processes and

$$v_{\alpha_1 1 \alpha_2}(u) = \sum_{n=0}^{\infty} P^n_{\alpha_2 G}(c_{\alpha_2} - g)(u) + (Q_{\alpha_2} v_{\alpha_1})(u) \quad , u \in G$$

$$v_{\alpha_1 1 \alpha_2}(u) = \sum_{n=0}^{\infty} P^n_{\alpha_1 F}(c_{\alpha_1} - g)(u) + (Q_{\alpha_1} v_{\alpha_1 1 \alpha_2})(u), \ u \in F$$

and for n = 2, 3, 4, ...

$$v_{\alpha_1 n \alpha_2}(u) = \sum_{n=0}^{\infty} P^n_{\alpha_2 G}(c_{\alpha_2} - g)(u) + (Q_{\alpha_2} v_{\alpha_1 n-1 \alpha_2})(u), \ u \in G$$

$$v_{\alpha_1 n \alpha_2}(u) = \sum_{n=0}^{\infty} P^n_{\alpha_1 F}(c_{\alpha_1} - g)(u) + (Q_{\alpha_1} v_{\alpha_1 n \alpha_2})(u) \quad , u \in F$$

If $\pi_{\alpha_0}(F) = 0$ the sum $\sum_{n=0}^{\infty} P^n_{\alpha_2 G}(c_{\alpha_2} - g)(u)$ in these expressions has to be replaced by $\sum_{n=0}^{\infty} P^n_{\alpha_2 G'}(c_{\alpha_2} - g)(u) + Q'_E v_{\alpha_2}$, where $E \subset G$ is a maximal invariant set of P_{α_2}, $G' := G \backslash E$ and Q' is the embedded Markov process of P_{α_2} on $F \cup E$. Notice that $Q'_F = Q_{\alpha_2}$.

If $\pi_{\alpha_0}(G) = 0$ the sum $\sum_{n=0}^{\infty} P^n_{\alpha_1 F}(c_{\alpha_1} - g)(u)$ has to be replaced in the same way.

But in each of these cases ($\pi_{\alpha_0}(F) > 0$, $\pi_{\alpha_0}(G) > 0$; $\pi_{\alpha_0}(F) = 0$, $\pi_{\alpha_0}(G) = 1$; $\pi_{\alpha_0}(F) = 1$, $\pi_{\alpha_0}(G) = 0$) it is easy to verify that.

$$\min\{v_{\alpha_1}(u), v_{\alpha_2}(u)\} - v_{\alpha_1 n \alpha_2}(u) \geq \sum_{\ell=1}^{n} R^{\ell}_{\alpha_0}(v_{\alpha_2} - v_{\alpha_1})(u) \quad , u \in V \qquad (*)$$

Let g_{α_0} be the average costs of the strategy α_0.

Using $v_{\alpha_0} = c_{\alpha_0} - g_{\alpha_0} + P_{\alpha_0} v_{\alpha_0}$ we get, for the case that $\pi_{\alpha_0}(F) > 0$, $\pi_{\alpha_0}(G) > 0$,

$$v_{\alpha_0}(u) = \sum_{n=0}^{\infty} P^n_{\alpha_2 G}(c_{\alpha_2} - g_{\alpha_0})(u) + (Q_{\alpha_2} v_{\alpha_0})(u) \ , \ u \in G$$

$$v_{\alpha_0}(u) = \sum_{n=0}^{\infty} P^n_{\alpha_1 F}(c_{\alpha_1} - g_{\alpha_0})(u) + (Q_{\alpha_1} v_{\alpha_0})(u), \ u \in F$$

If $g_{\alpha_0} = g$ then $v_{\alpha_1 n \alpha_2} = v_{\alpha_0} + R^n_{\alpha_0}(v_{\alpha_1} - v_{\alpha_0})$ and if $g_{\alpha_0} > g$ then $v_{\alpha_1 n \alpha_2} \to +\infty$ for $n \to \infty$, but this is impossible by (*) since $\sum_{\ell=1}^{n} R_{\alpha_0}(v_{\alpha_2} - v_{\alpha_1}) \geq 0$. Hence $g_{\alpha_0} = g$ and $v_{\alpha_1 n \alpha_2} = v_{\alpha_0} + R^n_{\alpha_0}(v_{\alpha_1} - v_{\alpha_0})$. This holds also for the cases $\pi_{\alpha_0}(F) = 1$, $\pi_{\alpha_0}(G) = 0$ and $\pi_{\alpha_0}(F) = 0$, $\pi_{\alpha_0}(G) = 1$.

Therefore

$$\min\{v_{\alpha_1}(u), v_{\alpha_2}(u)\} - v_{\alpha_0}(u) - R^n_{\alpha_0}(v_{\alpha_1} - v_{\alpha_0})(u) \geq \sum_{\ell=1}^{n} R^{\ell}_{\alpha_0}(v_{\alpha_2} - v_{\alpha_1})(u)$$

The boundedness of the sequence $R^n_{\alpha_0}(v_{\alpha_1} - v_{\alpha_0})(u)$ in n implies the convergence of the sum $\sum_1^{\infty} R^{\ell}_{\alpha_0}(v_{\alpha_2} - v_{\alpha_1})(u)$. But since $v_{\alpha_2} > v_{\alpha_1}$ everywhere on F this implies that the entry process R_{α_0} is absorbing, that means $\pi_{\alpha_0}(F) = 0$ or $\pi_{\alpha_0}(G) = 0$.

Hence $R^n_{\alpha_0}(v_{\alpha_1} - v_{\alpha_0})(u) \to 0$ and

$$v_{\alpha_0}(u) \leq \min\{v_{\alpha_1}(u), v_{\alpha_2}(u)\} - \sum_{\ell=1}^{\infty} R^{\ell}_{\alpha_0}(v_{\alpha_2} - v_{\alpha_1})(u)$$

This completes the proof of <u>ii</u> for the case that P_{α_0} has only one ergodic set. If P_{α_0} has more disjoint ergodic sets the proof can be given in the same way by considering the process on each of these sets. □

2. Existence of average optimal and sensitive optimal strategies

In this section an SMD $\{(P_\alpha, c_\alpha)\}$, $\alpha \in A$ is considered such that

<u>i</u> P_α is quasi-compact for all $\alpha \in A$

<u>ii</u> c_α is bounded on V, uniform in α

<u>iii</u> A is a metric space, metric ρ, such that

$$\lim_{\rho(\alpha,\alpha_0)\to 0} ||P_\alpha - P_{\alpha_0}|| \to 0 \text{ for all } \alpha_0 \in A$$

$$\lim_{\rho(\alpha,\alpha_0)\to 0} ||c_\alpha - c_{\alpha_0}|| \to 0 \text{ for all } \alpha_0 \in A$$

Let g_α, v_α be the average costs and the bias of (P_α, c_α). The strategy $\alpha_0 \in A$ is called <u>sensitive optimal</u> if α_0 is average optimal and if $v_{\alpha_0}(u) \le v_\alpha(u)$ for all $u \in V$ and all average optimal strategies α.
We will derive conditions for the existence of sensitive optimal strategies using the compactness of A and the continuity of P_α and c_α.
Define A_n, $n = 1, 2, \ldots$ as the set of all $\alpha \in A$ such that P_α has n disjoint ergodic sets. In the following lemma the continuity of g_α and v_α on A_n is stated. The proof is analogous to the proof of lemma 1.15 in [12] and uses operator valued functions and perturbation theory of linear operators (see Dunford-Schwartz [3], VII)

<u>Lemma 2</u>. Let $\{\alpha_i\}$ be a sequence in A_n converging to $\alpha_0 \in A_n$. Then $\lim_{i\to\infty} ||g_{\alpha_0} - g_{\alpha_i}|| = 0$ and $\lim_{i\to\infty} ||v_{\alpha_0} - v_{\alpha_i}|| = 0$

The following example shows that the continuity of v_α does not hold on the whole space A.

<u>Example</u>: Let $\{(P_\alpha, c_\alpha)\}$, $\alpha \in A$ be a problem with two states given by

$$P_\alpha = \begin{pmatrix} 1-\alpha & \alpha \\ 0 & 1 \end{pmatrix}, \quad c_\alpha = \begin{pmatrix} -\sqrt{\alpha} \\ 0 \end{pmatrix}, \quad A = \{\alpha | 0 \le \alpha \le \tfrac{1}{2}\}$$

Then $g_\alpha = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ for all $\alpha \in [0, \frac{1}{2}]$,

$$v_\alpha = \begin{pmatrix} -\frac{\sqrt{\alpha}}{\alpha} \\ 0 \end{pmatrix} \quad \text{for } \alpha > 0$$

and $v_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

Hence $v_\alpha(1)$ has a discontinuity in $\alpha = 0$. This discontinuity is due to the fact that for $\alpha > 0$ there is only one ergodic set and for $\alpha = 0$ two.

If in general $\{\alpha_\ell\}$ is a sequence in A_1 converging to $\alpha_0 \in A_n$ then in each neighbourhood of 1 (in the complex plane) there are eigenvalues of P_{α_ℓ} for ℓ large enough. Assume that the spectrum of the operators P_{α_ℓ} is of the following structure, $\sigma(P_{\alpha_\ell}) = 1 \cup \{\lambda_\ell\} \cup \sigma_\ell$ where $\lambda_\ell \to 1$ for $\ell \to \infty$ and σ_ℓ is for all ℓ a set within a circle with radius $\rho < 1$ (ρ independent of ℓ).
Let g_{λ_ℓ} be the projection of $c_{\alpha_\ell} - g_{\alpha_\ell}$ on $N((\lambda_\ell - P_{\alpha_\ell})^{\nu_\ell})$, where ν_ℓ is the index of λ_ℓ as eigenvalue of P_{α_ℓ}. Then

$$\lim_{\ell\to\infty} (v_{\alpha_\ell} - \frac{1}{1-\lambda_\ell} g_{\alpha_\ell}) = v_{\alpha_0} \text{ and } \lim_{\ell\to\infty}(g_{\lambda_\ell} + g_{\alpha_\ell}) = g_{\alpha_0}$$

In the example $g_{\lambda_\ell} = -\sqrt{\alpha_\ell}$, $\lambda_\ell = 1 - \alpha_\ell$

Remark. The average costs g_α have as function of α the same sort of discontinuities, but it is possible to define a rather general class of problems (communicating systems) where the set of all strategies A is dominated by the set of all strategies with a unique invariant probability. The communicativeness is introduced by Bather [1] for a finite state space and used by Hordijk [5] for a countable state space and Wijngaard [12] for a general state space.

To investigate the existence of sensitive optimal strategies we have to consider first the existence of average optimal strategies. This is done in the next theorem.

Theorem 3. Let A be compact, A_n closed in A for all $n = 1, 2, 3, \ldots$ and the number of ergodic sets of P_α bounded in α. Assume that the product property is satisfied. Then an average optimal strategy exists.

Proof. From lemma 2 and the assumption it follows immediately that for each $u \in V$ there is a strategy $\alpha_u \in A$ such that $g_{\alpha_u}(u) \le g_\alpha(u)$ for all $u \in V$ and all $\alpha \in A$ (the strategy α_u is u-optimal). Since A is a compact metric space it is separable. Let $\{\alpha_n\}_1^\infty$ be a countable subset of A which is dense in A. Then $\inf_n g_{\alpha_n}(u) = g_{\alpha_u}(u)$ for all $u \in V$. Let the strategies γ_n, $n = 1, 2, \ldots$ be such that $g_{\gamma_1} = g_{\alpha_1}$ and $g_{\gamma_n} \le \min\{g_{\gamma_{n-1}}, g_{\alpha_n}\}$ for all $n = 2, 3, 4, \ldots$. The existence of such strategies g_{γ_n} is guaranteed by lemma 1. The sequence $g_{\gamma_n}(u)$ is then monotonically non-increasing for each $u \in V$ and $g_{\gamma_n}(u) \leq g_{\alpha_n}(u)$. Hence $\lim_{n\to\infty} g_{\gamma_n}(u) = g_{\alpha_u}(u)$, $u \in V$. The boundedness of the number of ergodic sets, the compactness of A and the closedness of A_n for each n implies the existence of an integer ℓ and a subsequence $\{\gamma_n\}$ in A_ℓ converging to some γ in A_ℓ.This strategy γ is average optimal. □

A condition for closedness of A_n for all n = 1, 2, 3, ... is given in the next lemma. For the proof we refer to [12].

<u>Lemma 4</u>. If there is a ρ, $0 < \rho < 1$ such that for all $\alpha \in A$ the spectrum of P_α has no points λ with $\rho < |\lambda| < 1$, then A_n is closed in A for all $n = 1, 2, 3, \ldots$.

If the conditions of theorem 3 are satisfied the existence of a sensitive optimal strategy can be proved in the same way as the existence of an average optimal strategy. The continuity of g_α in α implies the closedness and hence compactness of the set of all average optimal strategies. We have the following result.

<u>Theorem 5</u>. If the conditions of theorem 3 are satisfied, a sensitive optimal strategy exists.

If α_0 is a sensitive optimal strategy, it is easy to prove that

$$v_{\alpha_0} = \min_{\alpha\in A}\{c_\alpha - g + P_\alpha v_{\alpha_0}\}$$

, where A' is the set of all α such that $P_\alpha g = g$. But even in the finite state space the converse is not true (see Blackwell [2]). That means that the sensitive optimal strategy cannot be approximated in general by policy improvement. If successive approximations can be applied depends on the question if $V_n - ng$ converges to v_{α_0} (V_n are the minimal expected n-period costs). For a treatment of this problem, see for instance Hordijk, Schweitzer, Tijms [6], Tijms [9] and Federgruen, Schweitzer [4].

References

[1] Bather, J. (1973): *"Optimal decision procedures for finite Markov chains, , part II: Communicating systems"*. Adv. in Appl. Prob. 5, 521-540.

[2] Blackwell, D. (1962): *"Discrete dynamic programming"*. Ann. Math. Statist. 33 , 719-729.

[3] Dunford, N., Schwartz, J.T. (1958): *"Linear Operators, part I"*. Interscience publishers, New York.

[4] Federgruen, A., Schweitzer, P.J. (1976): *"Asymptotic behaviour of undiscounted value iteration in Markov decision problems"*. Report BW 44/76, Math. Centre, Amsterdam.

[5] Hordijk, A. (1974): *"Dynamic programming and Markov potential theory"*. Math. Centre Tracts, no. 51, Amsterdam.

[6] Hordijk, A., Schweitzer, P.J., Tijms, H. (1975): *"The asymptotic behaviour of the minimal total expected costs for the denumerable state Markovian decision model"*. Jnl. Appl. Prob. 12, 298-305.

[7] Hordijk, A., Sladky, K. (1975): *"Sensitive optimality criteria in countable state dynamic programming"*. Report BW 48/75, Math. Centre, Amsterdam.

[8] Miller, B.L., Veinott, A.F., Jr. (1969): *"Discrete dynamic programming with a small interest rate"*. Ann. Math. Statist. 40, 366-370.

[9] Tijms, H. (1975): *On dynamic programming with arbitrary state space, compact action space and the average return as criterion"*. Report BW 55/75, Math. Centre, Amsterdam.

[10] Veinott, A.F., Jr. (1966): *"On finding optimal policies in discrete dynamic programming with on-discounting"*. Ann. Math. Statist. 37, 1284-1294.

[11] Veinott, A.F., Jr. (1969): *"Discrete dynamic programming with sensitive discount optimality criteria"*. Ann. Math. Statist. 40, 1635-1660.

[12] Wijngaard, J. (1975): *"Stationary Markovian decision problems, discrete time, general state space"*. Dissertation, Eindhoven University of Technology.