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 l-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 \((V, \Sigma)\) 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 \[|f| := \sup_{u \in V} |f(u)| \text{ for all } f \in B(V, \Sigma),\] then \(|.|\) is a norm on \(B(V, \Sigma)\) and \(B(V, \Sigma)\) is a Banach space.

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

\[(Pf)(u) = \int f(v)P(u, dv), \quad f \in B(V, \Sigma)\]

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

$$Q(u, E) = \sum_{n=0}^{\infty} \prod_{i=0}^{n-1} P_B(u, E) P_A(u, E)$$

If $\lim_{n \to \infty} (P^n_B |_V)(u) = 0$ 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 Markov process with costs. If $P$ is quasi-compact (satisfies the Doeblin condition) and $c$ is bounded, the average costs $g := \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k c$ exist and the functions $w := \lim_{m \to \infty} \frac{1}{d} \sum_{k=0}^{d-1} P^k (c-g)$ exist for all $m = 0, 1, 2, \ldots$ 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 up to a constant. The function $v$ is called the bias of $(P, c)$.

A stationary Markovian decision problem (SMD) is a set of Markov processes with costs $\{(P_\alpha, c_\alpha)\}, \alpha \in A$. The elements $\alpha \in A$ are called strategies. 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 product property.

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

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

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

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 \( P_\alpha \) quasi-compact and \( c_\alpha \) bounded on \( V \), uniform in \( \alpha \). 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

1. there exists a strategy \( \alpha_0 \in A \) such that

\[
g_{\alpha_0}(u) \leq \min\{g_{\alpha_1}(u), g_{\alpha_2}(u)\} \quad \text{for all } u \in V
\]

2. if \( \alpha_1, \alpha_2 \) are both average optimal then there exists a strategy \( \alpha_0 \in A \) such that

\[
v_{\alpha_0}(u) \leq \min\{v_{\alpha_1}(u), v_{\alpha_2}(u)\} \quad \text{for all } u \in V
\]

**Proof.** For the proof of the first part we refer to [12], section 4.1.3. Now let \( \alpha_1, \alpha_2 \) 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 \( Q_{\alpha_1} \) be the embedded sub-Markov process of \( P_{\alpha_1} \) on \( F \) and \( Q_{\alpha_2} \) the embedded sub-Markov process of \( P_{\alpha_2} \) on \( G \).

The strategy \( \alpha_0 \) is chosen such that

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

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

The product property implies that there is such a strategy \( \alpha_0 \) in \( A \). Let \( R_{\alpha_0} \) be the entry process of \( P_{\alpha_0} \) on \( F \), that means that \( R_{\alpha_0} \) is the sub-Markov process which describes the state of the system each time the set \( F \) is entered,

\[
R_{\alpha_0}(u, E) = Q_{\alpha_2}(u, E) \quad , \quad u \in G
\]

\[
R_{\alpha_0}(u, E) = (Q_{\alpha_1} Q_{\alpha_2})(u, E) \quad , \quad u \in F
\]
Define $v_{a_0^n a_2}$ as the bias of the (non-stationary) strategy which applies $a_0$ until the set $F$ is entered for the $n^{th}$ time and from then on the strategy $a_1$.

Consider first the case that $a_0$ has only one invariant probability $\pi_{a_0}$.

If $\pi_{a_0}(F) > 0$ and $\pi_{a_0}(G) > 0$ then $Q_{a_2}$ and $Q_{a_1}$ are Markov processes and

$$v_{a_1}a_2(u) = \sum_{n=0}^{\infty} \alpha_2 G_{a_2} \left( c_{a_2} - g \right)(u) + (Q_{a_2} v_{a_1})(u), \ u \in G$$

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

$$v_{a_1}a_2(u) = \sum_{n=0}^{\infty} \alpha_2 F_{a_2} \left( c_{a_2} - g \right)(u) + (Q_{a_2} v_{a_1})(u), \ u \in F$$

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

If $\pi_{a_0}(F) = 0$ the sum $\sum_{n=0}^{\infty} \alpha_2 G_{a_2} \left( c_{a_2} - g \right)(u)$ in these expressions has to be replaced by $\sum_{n=0}^{\infty} \alpha_2 G_{a_2} \left( c_{a_2} - g \right)(u) + Q' v_{a_1}$, where $E \subseteq G$ is a maximal invariant set of $P_{a_2}$, $G' := G \setminus E$ and $Q'$ is the embedded Markov process of $P_{a_2}$ on $F \cup E$. Notice that $Q' = Q_{a_2}$.

If $\pi_{a_0}(G) = 0$ the sum $\sum_{n=0}^{\infty} \alpha_1 F_{a_1} \left( c_{a_1} - g \right)(u)$ has to be replaced in the same way.

But in each of these cases ($\pi_{a_0}(F) > 0$, $\pi_{a_0}(G) > 0$; $\pi_{a_0}(F) = 0$, $\pi_{a_0}(G) = 1$; $\pi_{a_0}(F) = 1$, $\pi_{a_0}(G) = 0$) it is easy to verify that.

$$\min \{ v_{a_1}(u), v_{a_2}(u) \} - v_{a_1 a_2}(u) \geq \sum_{\ell=1}^{n} R_{a_0}^{\ell} \left( v_{a_2} - v_{a_1} \right)(u), \ u \in V \quad (*)$$

Let $g_{a_0}$ be the average costs of the strategy $a_0$.

Using $v_{a_0} = c_{a_0} - g_{a_0} + P_{a_0} v_{a_0}$ we get, for the case that $\pi_{a_0}(F) > 0$, $\pi_{a_0}(G) > 0$, $\pi_{a_0}(F) = 0$, $\pi_{a_0}(G) = 1$, $\pi_{a_0}(F) = 1$, $\pi_{a_0}(G) = 0$.
\[ v_{a_0}^0(u) = \sum_{n=0}^{\infty} p^n_{a_2 a_0} (c_{a_2} - g_{a_0})(u) + (Q_{a_2} v_{a_0})(u), \quad u \in G \]

\[ v_{a_0}^n(u) = \sum_{n=0}^{\infty} p^n_{a_1 a_0} (c_{a_1} - g_{a_0})(u) + (Q_{a_1} v_{a_0})(u), \quad u \in F \]

If \( g_{a_0} = g \) then \( v_{a_1 a_2 a_0} = v_{a_0} + R^n_{a_0 a_1 a_2}(v_{a_1} - v_{a_0}) \) and if \( g_{a_0} > g \) then

\[ \sum_{n=1}^{\infty} R^n_{a_0 a_1 a_2}(v_{a_1} - v_{a_0}) \geq 0. \] Hence \( g_{a_0} = g \) and \( v_{a_1 a_2 a_0} = v_{a_0} + R^n_{a_0 a_1 a_2}(v_{a_1} - v_{a_0}) \). This holds also for the cases \( \pi_{a_0}(F) = 1, \pi_{a_0}(G) = 0 \) and \( \pi_{a_0}(F) = 0, \pi_{a_0}(G) = 1 \).

Therefore

\[ \min\{v_{a_1}^n(u), v_{a_2}^n(u)\} - v_{a_0}^n(u) - R^n_{a_0 a_1 a_2}(v_{a_1} - v_{a_0})(u) \geq \sum_{k=1}^{\infty} R^k_{a_0 a_1 a_2}(v_{a_1} - v_{a_0})(u) \]

The boundedness of the sequence \( R^n_{a_0 a_1 a_2} v_{a_1} - v_{a_0} \) in \( n \) implies the convergence of the sum \( \sum_{k=1}^{\infty} R^k_{a_0 a_1 a_2} v_{a_1} - v_{a_0} \) in \( F \). But since \( v_{a_2} > v_{a_1} \) everywhere on \( F \) this implies that the entry process \( R_{a_0 a_1 a_2} \) is absorbing, that means \( \pi_{a_0}(F) = 0 \) or \( \pi_{a_0}(G) = 0 \).

Hence \( R^n_{a_0 a_1 a_2} v_{a_1} - v_{a_0} \) \( u \to 0 \) and

\[ v_{a_0}^n(u) \leq \min\{v_{a_1}^n(u), v_{a_2}^n(u)\} - \sum_{k=1}^{\infty} R^k_{a_0 a_1 a_2} v_{a_1} - v_{a_0} \]

This completes the proof of \( \text{ii} \) for the case that \( P_{a_00} \) has only one ergodic set.

If \( P_{a_0} \) has more disjoint ergodic sets the proof can be given in the same way by considering the process on each of these sets.

\[ \boxed{\text{2. Existence of average optimal and sensitive optimal strategies}} \]

In this section an SMD \( (P_{a_0}, c_{a_0}), \alpha \in A \) is considered such that

\[ \text{i} \quad P_{a_0} \text{ is quasi-compact for all } \alpha \in A \]

\[ \text{ii} \quad c_{a_0} \text{ is bounded on } V, \text{ uniform in } \alpha \]

\[ \text{iii} \quad A \text{ is a metric space, metric } \rho, \text{ such that} \]

\[ \boxed{\text{existence of average optimal and sensitive optimal strategies}} \]
Let $g_{\alpha}, v_{\alpha}$ be the average costs and the bias of $(P_{\alpha}, c_{\alpha})$. The strategy $\alpha_0 \in A$ is called sensitive optimal if $\alpha_0$ is average optimal and if $v_{\alpha_0}(u) \leq v_{\alpha}(u)$ for all $u \in V$ and all average optimal strategies $\alpha$.

We will derive conditions for the existence of sensitive optimal strategies using the compactness of $A$ and the continuity of $P_{\alpha}$ and $c_{\alpha}$.

Define $A_n$, $n = 1, 2, \ldots$ as the set of all $\alpha \in A$ such that $P_{\alpha}$ has $n$ disjoint ergodic sets. In the following lemma the continuity of $g_{\alpha}$ and $v_{\alpha}$ 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).

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

The following example shows that the continuity of $v_{\alpha}$ does not hold on the whole space $A$.

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

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

Then $g_{\alpha} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ for all $\alpha \in [0, \frac{1}{2}]$, $v_{\alpha} = \begin{pmatrix} -\sqrt{\alpha} \\ \frac{1}{\alpha} \end{pmatrix}$ 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 \( \{a_n\} \) is a sequence in \( A \) converging to \( a_0 \in A \), then in each neighbourhood of \( l \) (in the complex plane) there are eigenvalues of \( P \) for \( l \) large enough. Assume that the spectrum of the operators \( P \) is of the following structure, \( \sigma(P) = 1 \cup (\lambda_k) \cup (\sigma_k) \) where \( \lambda_k \to 1 \) for \( k \to \infty \) and \( \sigma_k \) is for all \( k \) a set within a circle with radius \( \rho < 1 \) (\( \rho \) independent of \( k \)).

Let \( g_{\lambda_k} \) be the projection of \( c - g \) on \( N((\lambda_k - P)^{-1}) \), where \( \nu_k \) is the index of \( \lambda_k \) as eigenvalue of \( P \). Then

\[
\lim_{k \to \infty} (\nu_k - \frac{1}{1-\lambda_k} g_{\lambda_k}) = \nu_0 \quad \text{and} \quad \lim_{k \to \infty} (g_{\lambda_k} + g_{\sigma_k}) = g_{\sigma_0}
\]

In the example \( g_{\lambda_k} = -\sqrt{\lambda_k}, \lambda_k = 1 - \alpha_k \)

Remark. The average costs \( g_{\alpha} \) have as function of \( \alpha \) 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 \( P \) bounded in \( \alpha \). 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 \( g_{\alpha_u}(u) \leq g_{\alpha}(u) \) for all \( u \in V \) and all \( \alpha \in A \) (the strategy \( \alpha_u \) is \( u \)-optimal). Since \( A \) is a compact metric space it is separable. Let \( \{\alpha_n\}_{n=1}^{\infty} \) be a countable subset of \( A \) which is dense in \( A \). Then \( \inf g_{\alpha_n}(u) = g_{\alpha}(u) \) for all \( u \in V \). Let the strategies \( \gamma_n \), \( n = 1, 2, \ldots \) be such that \( g_{\gamma_1} = g_{\alpha_1} \) and \( g_{\alpha} \leq \min\{g_{\gamma_{n-1}}, g_{\alpha_n}\} \) for all \( n = 2, 3, 4, \ldots \). The existence of such strategies \( g_{\gamma_n} \) is guaranteed.
by lemma 1. The sequence \( g_n(u) \) is then monotonically non-increasing for each \( u \in V \) and \( g_n(u) \leq g_\alpha(u) \). Hence \( \lim_{n \to \infty} g_n(u) = g_\alpha(u) \), \( u \in V \). 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 \( k \) and a subsequence \( \{ \gamma_n \} \) in \( A_k \) converging to some \( \gamma \) in \( A_k \). This strategy \( \gamma \) is average optimal.

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

Lemma 4. If there is a \( \rho, 0 < \rho < 1 \) such that for all \( \alpha \in A \) the spectrum of \( P_\alpha \) has no points \( \lambda \) 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 \( g_\alpha \) in \( \alpha \) implies the closedness and hence compactness of the set of all average optimal strategies. We have the following result.

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

If \( \alpha_0 \) is a sensitive optimal strategy, it is easy to prove that

\[
\nu = \min_{\alpha' \in A'} (c - g + P \nu),
\]

where \( A' \) is the set of all \( \alpha \) 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 \( V_n - n g \) converges to \( \nu_0 \) (\( V_n \) 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


