Markov decision processes with unknown transition law: the average return case
van Hee, K.M.

Published: 01/01/1979

Document Version
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Memorandum COSOR 79-01

Markov decision processes with unknown transition law; the average return case

by

Kees M. van Hee

Eindhoven, January 1979

The Netherlands
0. ABSTRACT

In this paper we consider some problems and results in the area of Markov decision processes with an incompletely known transition law. We concentrate here on the average return under the Bayes criterion. We discuss easy-to-handle strategies which are optimal under some conditions. For detailed proofs we refer to a monograph published by the present author.

1. INTRODUCTION

In this paper we review a part of (van Hee (1978a)), a monograph dealing with Markov decision processes in discrete time, with an incompletely known transition law. All proofs of statements given here, can be found in this monograph. Moreover this monograph contains results for the discounted return case; some of these results are reviewed in (van Hee (1978b)).

We do not bother about measure theoretic problems and therefore we assume all sets to be countable or sometimes even finite, however we remark that in (van Hee (1978a)) the problems are treated in a general measure theoretic setting.

We start with a description of the model and we discuss some of its properties. A Markov decision process (MDP) with unknown transition law is specified by a 5-tuple

\[(X, A, \Theta, P, r)\]  

where \(X\) is the state space, \(A\) the action space, \(\Theta\) the parameter space, \(P\) a transition probability from \(X \times A \times \Theta\) to \(X\) and \(r\) the reward function (i.e. \(r: X \times A \rightarrow \mathbb{R}\), where \(\mathbb{R}\) is the set of real numbers). (We assume \(r\) to be bounded, if \(X\) or \(A\) is countable). The parameter \(\Theta \in \Theta\) is unknown to the decision maker.

At each stage 0, 1, 2,... the decision maker chooses an action \(a \in A\) where he may base his choice on the sequence of past states and actions.

A strategy \(\pi\) is a sequence \(\pi = (\pi_0, \pi_1, \pi_2, \ldots)\) where \(\pi_0\) is a transition probability from \(X\) to \(A\) and \(\pi_n\) a transition probability from \((X \times A)^n \times X\) to \(A\) \((n \geq 1)\).

The set of all strategies is denoted by \(\Pi\).
A strategy is called stationary if there is a function \( f : X \rightarrow A \) such that always

\[
\pi_n \left( \{ f(x_n) \} \mid x_0, a_0, x_1, a_1, \ldots, x_n \right) = 1.
\]

According to the well-known Ionescu Tulcea theorem \cite{Neveu1965} we have for each starting state \( x \in X \), each strategy \( \pi \in \Pi \) and each parameter \( \theta \in \Theta \) a probability \( P_{x, \theta}^\pi \) on

\[
\Omega := (X \times A)^\mathbb{N} \quad (\mathbb{N} := \{0, 1, 2, \ldots \})
\]

and a random process \( \{(X_n, A_n), n \in \mathbb{N}\} \) where

\[
X_n(\omega) = x_n, \quad A_n(\omega) = a_n \quad \text{if } \omega = (x_0, a_0, x_1, a_1, \ldots) \in \Omega
\]

(The expectation with respect to \( P_{x, \theta}^\pi \) is denoted by \( E_{x, \theta}^\pi \).)

The average return is defined by

\[
g(x, \theta, \pi) := \liminf_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} E_{x, \theta}^\pi \left( r(X_n, A_n) \right)
\]

It only happens in non-interesting cases that there is a strategy \( \pi' \in \Pi \) such that \( g(x, \theta, \pi') \geq g(x, \theta, \pi) \) for all \( x \in X \) and \( \theta \in \Theta \).

So we cannot use this as an optimality criterion. We have chosen the Bayes criterion \cite{vanHee1978} for a motivation cf. \cite{vanHee1978a}.

Fix some probability \( q \) on \( \Theta \). (Such a probability is called a prior distribution.)

The Bayesian average return with respect to \( q \) is defined by

\[
g(x, q, \pi) := \liminf_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \sum_{\theta} q(\theta) E_{x, \theta}^\pi \left( r(X_n, A_n) \right)
\]

(Note that the definitions (1.4) and (1.5) are consistent if we identify \( \Theta \) with the distribution that is degenerate at \( \theta \).)

The set of all probabilities on \( \Theta \) will be denoted by \( W \). A strategy \( \pi' \) is called \( \varepsilon \)-optimal in \( (x, q) \in X \times W \) if

\[
g(x, q, \pi') \geq g(x, q, \pi) - \varepsilon \quad \text{for all } \pi \in \Pi
\]

(a 0-optimal strategy is simply called optimal).

The Bayes criterion allows us to consider the parameter as a random variable \( Z \) with range \( \Theta \) and distribution \( q \) on \( \Theta \). On \( \Theta \times \Omega \) we have the probability \( P_{x, \theta}^\pi \) determined by

\[
P_{x, q}^\pi \left( Z \in B, (X_0, A_0, X_1, A_1, \ldots) \in C \right) = \sum_{\theta \in B} P_{x, \theta}^\pi \left( C \right) q(\theta)
\]

for events \( C \) in \( \Omega \).
We compute the so-called posterior distributions $Q_n$ of $Z$ in the following way:

$$Q_n(B) := \mathbb{P}^\pi_{x,q} \left[ Z \in B \mid X_0, A_0, X_1, A_1, \ldots, X_n, A_n \right]$$

(Note that $Q_n$ is determined $\mathbb{P}^\pi_{x,q}$ a.s.)

Define the probability $T_{x,a,x'}(q)$ on $\Theta$ by

$$T_{x,a,x'}(q)(\theta) := \frac{P(x' \mid x,a,\theta) q(\theta)}{\sum_{\theta'} P(x' \mid x,a,\theta') q(\theta')}$$

It is possible to choose versions of the posterior distributions such that

$Q_0 = q$ and $Q_{n+1} = T_{X_n,A_n,X_{n+1}}(Q_n)$.

As indicated by Bellman (cf. Bellmann (1961)) and proved in a very general setting in Rieder (1975) this decision model is equivalent to a MDP with a known transition law, specified by a 4-tuple

$$(X \times W, A, P, R)$$

where $X \times W$ is state space, $A$ the action space, $P$ the transition law defined by

$$P(x', T_{x,a,x'}(q) | x,q,a) := \sum_{\theta} q(\theta) P(x' | x,a,\theta)$$

and $R : X \times W \times A \to \mathbb{R}$, the reward function, is defined by

$$R(x,q,a) := r(x,a)$$

Note that the state $(x,q)$ of the new model (1.10) consists of the original state $x \in X$ and the "information state" $q \in W$. It turns out that each state $(x,q)$ and each strategy $\pi$ for the new model, define a probability $P^\pi_{x,q}$ and a random process $\{X_n, Q_n, A_n, n \in \mathbb{N}\}$ on $\Omega := (X \times W \times A)^N$. Here is

$X_n(\omega) := x_n, Q_n(\omega) := q_n$ and $A_n(\omega) := a_n$ where $\omega = (x_0, q_0, a_0, x_1, q_1, a_1, \ldots) \in \Omega$.

The original model (1.1) and the new model (1.10) have the following relationship:

$$E_{x,q} \left[ r(X_n, A_n) \right] = E^\pi_{x,q} \left[ R(X_n, A_n) \right]$$

where $\pi$ is the strategy for model (1.10), which is defined by

$$\pi_n(a_n | x_0, q_0, a_0, \ldots, x_n, q_n) := \pi_n(a_n | x_0, a_0, \ldots, x_n)$$

Hence models (1.1) and (1.10) are equivalent and therefore we use the notations of model (1.1).
So we are dealing with a Markov decision process with known transition law again. However this new MDP has some odd properties. At first the state space is infinite even if the state space of the original model is finite. Further the new MDP is transient in general, i.e. in most cases \( Q_n \neq Q_m \).

\[ P_{x,q} \text{ a.s. for all } n \neq m. \]

In section 2 we show by an example that even if \( X \) and \( A \) are finite sets there need not to be an optimal strategy.

In the next section we introduce strategies that are easy-to-handle, at least if \( X, A \) and \( \Theta \) are finite sets, and we consider conditions guaranteeing these strategies to be optimal. These conditions imply that the posterior distributions \( Q_n \) converge to degenerate distributions, which property is used explicitly to prove the optimality.

We conclude this section by introducing a parameter structure that is quite general and that facilitates formulating some results.

From now on we assume that we are dealing with the following structure:

(i) \( X = \hat{X} \times Y \)

(ii) \( R \) is a transition probability from \( \hat{X} \times A \times Y \) to \( \hat{X} \)

(iii) \( K_1, K_2, K_3, \ldots \) is a partition of \( \hat{X} \times A \),

(iv) \( \Theta = \prod_{i=1}^{\infty} \Theta_i, \Theta = (\Theta_1, \Theta_2, \Theta_3, \ldots) \)

(v) \( P(x', y' \mid x, a, \Theta) = R(x' \mid x, a, y') \cdot p_i(y' \mid \Theta_i) \) if and only if \( (x, a) \in K_i \)

where \( p_i(. \mid \Theta_i) \) is a probability on \( Y \).

Hence we have factorized the original transition law. If \( (x, a) \in K_i \) then the transition to the next state \( x', y' \) depends on \( \Theta \) only through its \( i \)-th component \( \Theta_i \). We present below some examples having this structure. It is straightforward to verify that

\[
Q_{(x,y),\Theta, (x',y')} (q) (\Theta) = \sum_{i=1}^{\infty} 1_{K_i} (x,a) \frac{p_i(y' \mid \Theta_i) q(\Theta)}{\sum_{\Theta} p_i(y' \mid \Theta_i) q(\Theta)}
\]

(provided that the denominator does not vanish). Here \( 1_B \) represents the indicator function of the set \( B \).

Although this parameter structure seems to be rather complicated there are practical situations where this structure occurs in a natural way.
The state of the system at stage \( n \) is \( X_n = (\hat{X}_n, Y_n) \). The state component \( Y_n \) is called the *supplementary state variable*. It can be proved that if \( \theta \in \Theta \) is known then it is sufficient to consider \( \hat{X}_n \) instead of \( X_n \) (see [van Hee (1978a) page 52]).

So \( \hat{X}_n \) has to be considered as the original state variable if the parameter is known, while \( Y_n \) only occurs since it contains information concerning the unknown parameter.

From now on we are dealing with this parameter structure and we shall consider only \( \hat{X}_n \) and \( \hat{X} \). To facilitate notations we omit the head from now on.

**Example 1**

Consider an inventory control model without backlogging. If the demand distribution is known, the inventory level may be chosen as the state variable. However if the demand distribution is unknown the sequence of successive inventory levels does not reflect the sequence of successive demands and therefore we have to consider the demand in each period as a supplementary state variable. Here \( X \) is the set of inventory levels, i.e. \( X = (0, \infty) \) and \( Y \) is the set of possible demands. The transition function is

\[
R(x' | x, a, y') = \begin{cases} 
1 & \text{if } x' = \max \{a-y', 0\} \text{ and } a > x \\
0 & \text{otherwise}
\end{cases}
\]

(Hence the action \( a \) is the inventory level after ordering).

The sets \( K_1, K_2, \ldots \) are empty and \( p_1(\cdot | \theta) \) is the demand distribution with unknown parameter \( \theta \in \Theta \).

**Example 2**

Consider a waiting line model with bulk arrivals. At each time point \( 0, 1, 2, \ldots \) a group of customers arrives and the distribution of the size of the group is unknown. The service distribution is exponentially with parameter \( a \) and controllable by the parameter \( a \). Let \( y' \) be the number of customers arriving in some period, let \( x \) be the queue length at the beginning of that period and \( x' \) at the end. Then, if \( c: = x + y' - x' \geq 0 \) we have

\[
R(x' | x, a, y') = \frac{a^c}{c!} e^{-a}
\]
and if $c < 0$ then $R(x' | x, a, y') = 0$. Further $K_2, K_3, \ldots$ are empty and $p_1(y' | \theta_1)$ is the probability of a group if size $y'$.

**Example 3**

Consider a linear system with random disturbances. The state at stage $n$ is $X_n$ and the disturbance at stage $n$ is $Y_n$.

Then

$$X_{n+1} = C_1 X_n + C_2 A_n + Y_{n+1}$$

where $X, Y$ and $A$ are Euclidean spaces and $C_1$ and $C_2$ suitable matrices.

Assume that $\{Y_n, n \in \mathbb{N}\}$ forms a sequence of i.i.d. random variables with an incompletely known distribution.

If only the sequence $(X_0, X_1, X_2, \ldots)$ is observable to the controller then he may reconstruct the sequence of supplementary state variables.

2. **OPTIMAL STRATEGIES**

We start this section with an example showing that there need not to be an optimal strategy even if $X$ and $A$ are finite sets.

**Example 4**

Let $X = \{1,2,3,4,5,6\}$, $A = \{1,2,3\}$, $Y = \{0,1\}$, $\Theta_1 = \{0,1\}$.

(Note that $\Theta_1$ is not countable here but if we replace $\Theta_1$ by a countable subset of $\{0,1\}$ the same arguments are valid, however notations become more difficult).

Further let $R(x' | x, a, y')$ be defined by:

- $R(3|3,a,1) = R(4|4,a,1) = R(3|4,a,0) = R(4|4,a,0)$
- $R(5|5,a,1) = R(6|6,a,0) = R(5|6,a,1) = 1$ for all $a \in A$, and let
- $R(1|1,1,1) = R(2|2,1,0) = 1$; $R(3|1,2,y) = R(5|1,3,y) = 1$ for all $y \in Y$.

Finally $K_2, K_3, \ldots$ are empty and $p_1(1|\theta) = 0, p(0|\theta) = 1-0$. 
The example can be represented in the diagram:

Only in states 1 and 2 the chosen action has effect. The rewards obtained are \( r(3) = r(5) = 7 \) and \( r(4) = r(6) = 3 \) for all actions and in the other states no rewards are obtained.

The average return in the sub-chain \( \{3,4\} \) is \( \frac{1}{2} (7+3) = 5 \) and in the sub-chain \( \{5,6\} : 7 \theta + 3(1-\theta) = 4\theta + 3 \). Consider a starting state \( x \in \{1,2\} \).

It is easy to verify that for known \( \theta \in \Theta \), the optimal action is a maximizer of the function \( 5.\delta(2,a) + \{4\theta + 3\} \delta(3,a) \) (where \( \delta \) is the Kronecker function), \( a \in \{2,3\} \). It is also straightforward to verify that if we have to choose one of the actions 2 of 3 and if \( q \) is the prior distribution then the maximizer of \( 5.\delta(2,a) + \{4. \mathbb{I}q(\theta) + 3\} \delta(3,a) \), \( a \in \{2,3\} \) is the best choice.

Let \( \pi^n \) be the strategy that chooses action 1 the first \( n \) times and the maximizer of the function

\[
5\delta(2,a) + \{4 \mathbb{I} \theta Q_n(\theta) + 3\} \delta(3,a), a \in \{2,3\}
\]

thereafter, where \( Q_n \) is the posterior distribution at time \( n \) if the system starts in state 1 with prior distribution \( q \).

Then the Bayesian average return in states 1 and 2 is

\[
E_q \left[ \max \{5,4 \mathbb{I} \theta Q_n(\theta) + 3\} \right]
\]

(this expectation does not depend on the starting state and the strategy).

Note that

\[
E_q \left[ \max \{5,4 \mathbb{I} \theta Q_{n+1}(\theta) + 3 \mid Q_1, \ldots, Q_n\} \right] \geq \max \{5,4 E_q \left[ \mathbb{I} \theta Q_{n+1}(\theta) | Q_1, \ldots, Q_n\right] + 3\} = \max \{5,4 \mathbb{I} \theta Q_n(\theta) + 3\}
\]

with equality if and only if \( 5 \geq 4 \mathbb{I} \theta Q_{n+1}(\theta) + 3 \) \( P_q \) a.s. However if \( q \) gives positive mass to the set \( \{\theta \in \Theta \mid \theta > \frac{1}{2}\} \) the equality never holds. Hence in this case the strategy \( \pi^n \) is worse than \( \pi^{n+1} \) and consequently there is no optimal strategy.
We first introduce two assumptions:

1. \( r \) is bounded on \( X \times A \)  
2. there are bounded functions \( g \) and \( h \) on \( \Theta \), and \( X \times \Theta \) respectively, such that \( h(\theta, \Theta) + g(\theta) = \sup_{\alpha \in A} L(x, \alpha, \Theta) \)

where \( L(x, \alpha, \Theta) = \sum_{i=1}^{\infty} K_i(x, \alpha) \{ r(x, \alpha) + \sum \sum R(x' | x, \alpha, y') \cdot p_i(y' | \theta_i) \cdot h(x', \Theta) \} \)

For models with known parameter value these conditions are the well-known conditions considered in [Derman (1966)] and [Ross (1968)]. In that case there exist stationary optimal strategies, at least if \( A \) is finite. Moreover the optimal average return is \( g(\theta) \).

In [Ross (1968)] several situations are given where (2.1) (i.i) holds for fixed parameter value \( \theta \).

For instance, if \( X \) and \( A \) are finite sets, and for each stationary strategy the resulting process is an irreducible Markov chain, then (2.1) (i.i) is valid.

Let \( \{ \epsilon_n, n \in \mathbb{N} \} \) be a sequence of bounded real numbers such that \( \lim_{n \to \infty} \epsilon_n = 0 \).

The strategies we consider choose at each stage \( n \) a maximizer of

\[ \sum_{\theta} L(X_n, \alpha, \Theta) Q_n(\theta) , \alpha \in A \]  

We call these rules **Bayesian equivalent rules** because we are maximizing the "Bayesian equivalent" of the function we have to maximize in case the parameter is known. Note that we may choose \( \epsilon_n = 0 \), \( n \in \mathbb{N} \) if there is a maximizer of the function (2.2).

In example 4 we already encountered these strategies. So it is clear by the example that these strategies are not optimal in general. However we give at the end of this section conditions guaranteeing these strategies to be optimal.

In [van Hee (1978a)] we consider these strategies also for the discounted total return case and we prove there that these strategies are optimal for the linear system with quadratic costs (in discrete time) and also for some inventory control models. There we also consider bounds on the discounted total return of these strategies.
In [Fox and Rolph (1973)], [Mandl (1974)] and [Georgin (1978)] another heuristic strategy is considered, which turns out to be optimal in a lot of situations. This strategy can be formulated in the following way:

"At each stage estimate the unknown parameter \( \theta \) using the available data, by \( \hat{\theta} \). Then compute an optimal (stationary) strategy for the model where the parameter is known and equal to \( \hat{\theta} \). Then use the corresponding action in the actual state. Repeat this procedure at each stage."

Hence, if we consider Bayes estimates, then the method proposed by the other authors may be formulated in the following way:

choose at each stage \( n \) a maximizer of

\[
L(X_n, a, \Sigma \theta Q_n(\theta)), a \in A
\]

(2.3)

To prove the optimality of Bayesian equivalent rules we need the following limit theorem:

**Theorem 1**

If

\[
\sum_{n=0}^{\infty} 1_{K_i} (X_n, A_n) = \infty (P^\Pi_{x,q} - a.s.)
\]

then

\[
\lim_{n \to \infty} \Sigma f(\theta) Q_n(\theta) = f(Z_1) (P^\Pi_{x,q} - a.s.) i = 1,2,3,...
\]

for all bounded functions \( f \) on \( \Theta \).

(where \( Z = (Z_1, Z_2, Z_3, ...) \) cf (1.15) (i.v.)

Note that theorem 1 gives a sufficient condition for the consistency of the Bayes estimation procedure.

We introduce the function \( \phi \) on \( X \times A \times \Theta \).

\[
\phi(x, \theta, a) : = L(x, a, \theta) - h(x, \theta) - g(\theta)
\]

Note that \( \phi(x, \theta, a) \leq 0 \) for all \( x \in X, \theta \in \Theta, a \in A \).

Further we extend \( \phi \) to a function on \( X \times A \times W \) and likewise the functions \( h \) and \( g \).
(i) \[ \phi(x,q,a) := \sum_{\theta} q(\theta) \phi(x,\theta,a) \] 
(ii) \[ h(x,q) := \sum_{\theta} q(\theta) h(x,\theta) \] 
(iii) \[ g(q) := \sum_{\theta} q(\theta) g(\theta) \]

Note that these definitions are consistent of (2.1) (ii) if we identify \( \theta \) with the distribution that is degenerate at \( \theta \).

The next theorem provides us a sufficient condition for a strategy \( \pi \) to be optimal.

**Theorem 2**

If \[ \liminf_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} E_{x,q} [\phi(x_n,a_n)] = 0 \] Then \[ \liminf_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} E_{x,q} [r(x_n,a_n)] = g(q) \] and \( \pi \) is optimal.

Theorems 1 and 2 are the key tools for proving optimality of the Bayesian equivalent rules.

As an illustration we shall consider two practical examples from more general theorems. For the first example we provide a proof to demonstrate the technique.

**Theorem 3**

Let \( X \) and \( A \) be finite sets, let \( M_1, M_2, \ldots, M_m \) be a partition of \( X \) and let \( K_i = M_i \times A, i=1,\ldots,m (K_i=\emptyset \text{ for } i > m) \).

Let the Markov chain \( \{X_n, n \in \mathbb{N}\} \) be irreducible for each stationary strategy and each parameter value.

Then a strategy \( \pi^* \) that chooses at stage \( n \) a maximizer of the function

\[ \phi(X_n, Q_n, a), a \in A \]

is optimal.

(Note that a maximizer of \( \phi(X_n, Q_n, a) \) is a maximizer of \( \sum_{\theta} L(X_n,a,\theta).Q_n(\theta) \))

**Proof.**

Let \( A_n \) be the action at stage \( n \) under strategy \( \pi^* \). Hence

\[ 0 > \phi(X_n, Q_n, A_n) = \max_{a \in A} \sum_{\theta} Q_n(\theta) \phi(X_n, \theta, a) = \]

\[ \phi(X_n, Q_n, a) = \]
\[ \max \sum_{n} \min_{\theta} \phi(x, \theta, f(x)) = \max \sum_{n} \min_{\theta} \phi(x, \theta, f(x)) \]

where \( F \) is the set of all functions from \( X \) to \( A \).

Note that \( F \) represents the set of all stationary strategies for the model with known parameter value.

Since the Markov chain \( \{X_n, n \in \mathbb{N}\} \) is irreducible for each stationary strategy it can be proved (cf. lemma 4.7 in [van Hee (1978a)]) that the number of visits to each set \( M_i \) is almost surely infinite for all strategies.

Therefore the condition of theorem 1 is fulfilled for all \( i \) and so we have

\[ \lim_{n \to \infty} \sum_{\theta} \min_{x} \phi(x, \theta, f(x)) = \min_{x} \phi(x, \theta, f(x)) \]

Since there is a stationary optimal strategy for the model with known parameter value (cf. Ross (1968)) we have

\[ \max_{\theta} \min_{x} \phi(x, \theta, f(x)) = 0 \text{ for all } \theta \in \Theta. \]

Therefore we find

\[ 0 \geq \liminf_{n \to \infty} \phi(x_n, \theta, A_n) \geq \lim_{n \to \infty} \max_{\theta} \sum_{x} \phi(x_n, \theta, f(x)) = 0 \]

and so

\[ \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} E_{x, q}^\pi \left[ \phi(x_n, \theta, A_n) \right] = 0 \]

Application of theorem 2 gives the desired result.

In this example we assumed that the information we obtain after the transition about the unknown parameter does not depend on the action chosen. In the next example we relax this assumption.

Here we assume:

(i) \( M_1, \ldots, M_m \) is a partition of \( X \), \( N_1, \ldots, N_n \) is a partition of \( A \). (2.5)

(ii) For each stationary strategy the Markov chain \( \{X_t, t \in \mathbb{N}\} \) is irreducible

(iii) The partition \( K_1, K_2, \ldots \) of \( X \times A \) consists of the sets

\[ \{M_i \times N_j, i=1, \ldots, m, j=1, \ldots, n\}. \]
Before we consider the strategy that turns out to be optimal, we first introduce the concept of a sequence of density zero. A sequence $S = (s_1, s_2, \ldots)$ is said to be of density zero if

$$\limsup_{k \to \infty} \max \{i \in \mathbb{N} \mid s_i = k\} = 0$$

Examples of such sequences are: $(s_i = 2^i, i \in \mathbb{N})$ and $(s_i = i^2, i \in \mathbb{N})$

In theorem 4 we consider a strategy that is inspired by an idea in [Mallows and Robbins (1964)]. In [Fox and Rolph (1973)], this idea is used in a similar way for Markov renewal programs.

The idea is, that we make use of forced choice actions to guarantee that we return to each set $K_i$ infinitely often, which is necessary to apply theorem 1. However we do this with a frequency that is so low as not to influence the Bayesian average return. (In fact the concept of a sequence of density zero is used here).

Now we are ready to formulate in an informal way the strategy $\hat{\pi}$ that will be optimal. In (2.6) this strategy is sketched:

Fix (forced choice) actions $a_1, a_2, \ldots, a_n$ such that $a_i \in \mathbb{N}$ (2.6)

Let $t_i(n)$ be the number of visits to the set $M_i$ at stage $n$.

If $x_n \in M_i$ and $t_i(n) \in S$ for some $i \in \{1, \ldots, m\}$ then the next action in the sequence $(a_1, \ldots, a_n)$ is chosen.

Otherwise, if $t_{x_n}(n) \not\in S$, then a maximizer of the function

$$\phi(x_n, Q_n, a), a \in A$$

is chosen.

Hence $\hat{\pi}$ uses the same actions as the strategy $\pi^*$ in theorem 3 except for stages where $t_{x_n}(n) \in S$. Then the forced choice actions are chosen in order.

The proof of theorem 4 is rather technically, although the idea is simple.

Theorem 4

Let (2.5) hold. The strategy $\hat{\pi}$ defined in (2.6) is optimal.

This result is also true in a more abstract model. In fact it completes results of [Mandl (1974)] and it generalizes work of [Rose (1975)].
3. LITERATURE

Bellman, R., Adaptive control processes: a guided tour

Derman, C., Denumerable state Markovian decision processes -
average cost criterion,

Fox, B.L. and Rolph, J.E., Adaptive policies for Markov renewal programs

Georgin, J.P., Estimation et contrôles des chaînes de Markov sur des
espaces arbitraires.
In: Lecture Notes in Mathematics 636
Springer-Verlag, Berlin etc. (1978)

Van Hee, K.M., Bayesian control of Markov chains
Amsterdam, Mathematical Centre Tracts 95 (1978a)

Van Hee, K.M., Markov decisions processes with unknown transition law;
the discounted case
To appear in: Kybernetica (1978b)

Mallows, C.L. and Robbins, H., Some problems of optimal sampling strategy.

Mandl, P., Estimation and control in Markov chains
Advances in Appl. Probability 6 (1974), 40-60

Neveu, J., Mathematical foundations of the calculus of
probability
San Francisco etc., Holden-Day (1965)

Rieder, U., Bayesian dynamic programming
Advances in Appl. Probability 7 (1975), 330-348

Rose, J.S., Markov decision processes under uncertainty -
average return criterion
Unpublished report (1975)

Ross, S.M., Arbitrary state Markovian decision processes