Markov decision with unknown transition law: the discounted case

Citation for published version (APA):

Document status and date:
Published: 01/01/1978

Publisher 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 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

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.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Memorandum COSOR 78-16

Markov decision with unknown transition law; the discounted case

by

Kees M. van Hee

Eindhoven, september 1978
The Netherlands
Markov decision processes with unknown transition law; the discounted case

by

Kees M. van Hee

Abstract. In this paper we consider some problems and results in the field of Markov decision processes with an incompletely known transition law. We consider the discounted total return under the Bayes criterion. We discuss easy-to-handle strategies which are optimal under some conditions for the average return case and also for some special models in the discounted total return case. Further we provide approximation methods to compute the optimal value.

Introduction. In this paper we review a part of [van Hee (1978)], a monograph dealing with Markov decision processes with unknown transition law. All proofs of statements given here, can be found in this monograph. In this paper we do not bother about measure theoretic problems and therefore we assume all sets to be countable and sometimes even finite.

We start with a sketch of the problems and we give a quick overview of the contents of the following sections.

A Markov decision process (MDP) with unknown transition law is specified by a 5-tuple

$$\text{(X,A,θ,\mathbb{P},r)}$$

where X is the state space, A the action space, θ 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.

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 $\pi = (\pi_0, \pi_1, \pi_2, ...)$ 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 $\Pi$.

According to the well-known Ionescu Tulcea theorem (cf. [Neveu (1965)]) we have for each starting state $x \in X$, each strategy $\pi \in \Pi$ and each parameter $\theta \in \Theta$ a probability $\mathbb{P}^\pi_{x,\theta}$ 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 \( \mathbb{P}_{x, \theta}^{\pi} \) is denoted by \( \mathbb{E}_{x, \theta}^{\pi} \).

The most used values to rate strategies are the discounted total return

1.2. \( v(x, \theta, \pi) := \sum_{n=0}^{\infty} \beta^n \mathbb{E}_{x, \theta}^{\pi}[r(X_n, A_n)], \beta \in (0,1), x \in X, \theta \in \Theta, \pi \in \Pi \)

and the average return

1.3. \( g(x, \theta, \pi) := \liminf \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}_{x, \theta}^{\pi}[r(X_n, A_n)] \).

We only consider the value \( v(x, \theta, \pi) \) in this paper.

It seldom occurs that there is a strategy \( \pi^* \in \Pi \) such that

\[
v(x, \theta, \pi^*) \geq v(x, \theta, \pi) \quad \text{for all } \pi \in \Pi \text{ and all } \theta \in \Theta.
\]

So we need another criterion. We have chosen the Bayes criterion. Hence we fix a probability \( q \) on \( \Theta \) and we define

1.4. \( v(x, q, \pi) := \sum_{n=0}^{\infty} \beta^n \sum_{\theta \in \Theta} q(\theta) \mathbb{E}_{x, \theta}^{\pi}[r(X_n, A_n)] \)

and for the average return case

1.5. \( g(x, q, \pi) := \liminf \frac{1}{N} \sum_{n=0}^{N-1} \sum_{\theta \in \Theta} q(\theta) \mathbb{E}_{x, \theta}^{\pi}[r(X_n, A_n)] \).

(Note that these definitions are consistent with 1.2 and 1.3 if we identify \( \theta \in \Theta \) with the distribution that is degenerate at \( \theta \).

The set of all probabilities on \( \Theta \) is denoted by \( \mathcal{W} \). A strategy \( \pi^* \) is called \( \varepsilon \)-optimal in \( (x, q) \in X \times \mathcal{W} \) for the discounted total return case if

\[
v(x, q, \pi^*) \geq v(x, q, \pi) - \varepsilon \quad \text{for all } \pi \in \Pi.
\]

\( \varepsilon \)-optimal strategy is simply called optimal). Similarly for the average return case.

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 \( \mathbb{P}_{x, q}^{\pi} \) determined by

\[
\mathbb{E}_{x, q}^{\pi}[Z \in B, (X_0, A_0, X_1, A_1, \ldots) \in C] = \sum_{\theta \in B} q(\theta) \mathbb{E}_{x, \theta}^{\pi}[C].
\]

We may compute the so-called posterior distribution of \( Z \)

\[
Q_n(B) := \mathbb{E}_{x, q}^{\pi}[Z \in B \mid X_0, A_0, X_1, A_1, \ldots, X_n, A_n].
\]

(Note that \( Q_n \) is determined \( \mathbb{P}_{x, q}^{\pi} \)-a.s.)

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

1.6. \( T_{x, a, x'}(q)(\theta) := \frac{P(x' \mid x, a, \theta)q(\theta)}{\sum_{\theta} P(x' \mid x, a, \theta)q(\theta)}. \)
It is possible to choose versions of the posterior distributions such that $Q_0 = q$ and $Q_{n+1} = T_{X_n A_n X_n} (Q_n)$. As indicated by Bellman (cf. Bellman (1961)) and proved in a very general setting in [Rieder (1975)] the decision model is equivalent to a MDP with a known transition law, specified by a 4-tuple

$$\begin{align*}
&1.7. \quad (X \times W, A, \bar{P}, \bar{r})
\end{align*}$$

where $X \times W$ is the state space, $A$ the action space, $\bar{P}$ the transition probability, defined by

$$\begin{align*}
\bar{P}(x', T_{X_n A_n X_n}, Q_n) := \sum_{\theta} q(\theta) P(x' | x, a, \theta)
\end{align*}$$

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

$$\begin{align*}
\bar{r}(x, q, a) := r(x, a)
\end{align*}$$

Note that the state $(x, q)$ of the new model (1.7) consists of the original state $x \in X$ and the "information" state $q \in W$. It turns out that each starting state $(x, q) \in X \times W$ and each strategy $\pi$ for this model define a probability $\hat{P}_{x, q}$ and a random process 

$$\{(X_n, Q_n, A_n), n \in \mathbb{N}\} \text{ on } \hat{\Omega} := (X \times W \times A)^{\mathbb{N}}$$

Here $X_n (\omega) := x_n$, $Q_n (\omega) := q_n$ and $A_n (\omega) := a_n$ for $\omega = (x_0, q_0, a_0, x_1, q_1, a_1, \ldots) \in \hat{\Omega}$. The original model (1.1) and the new model (1.7) have the following relation for $n \in \mathbb{N}$:

$$\begin{align*}
1.9. \quad \mathbb{E}_{x, q}^n [r(X_n, A_n)] = \hat{E}_{x, q}^n [\bar{r}(X_n, Q_n, A_n)]
\end{align*}$$

where $\pi$ is the strategy for model 1.7 which is defined by

$$\begin{align*}
\pi_n (a_n | x_0, q_0, a_0, \ldots, x_n, q_n) := \pi_n (a_n | x_0, a_0, \ldots, x_n)
\end{align*}$$

Hence model 1.1 and model 1.7 are equivalent, and we shall use the notation for model 1.1. For model 1.7 we may apply all well-known theory for the determination of the value function $v$ on $X \times W$:

$$\begin{align*}
1.10. \quad v(x, q) := \sup_{\pi \in \Pi} v(x, q, \pi)
\end{align*}$$

($v(x, q, \pi)$ is defined in 1.4; we consider two functions $v$ with different domain). Unfortunately, even if $X$ and $A$ are finite, model 1.7 has a state space that is essentially infinite. Therefore we do not have algorithms to determine the function $(x, q) \rightarrow v(x, q)$.

However for fixed $q \in W$ we may compute $v(x, q)$, $x \in X$. To see this we introduce the well-known optimal reward operator $U$ for the MDP defined in 1.7 (we use notations of model 1.1)

$$\begin{align*}
1.11. \quad (U_b)(x, q) := \sup \{ r(x, a) + \beta \sum_{\theta} q(\theta) P(x' | x, a, \theta) b(x', T_{X_n A_n X_n}(q)) \}
\end{align*}$$

where $b: X \times W \rightarrow \mathbb{R}$ is bounded (and measurable).

The following properties hold:

$$\begin{align*}
1.12. \quad \lim_{n \rightarrow \infty} (U^n b)(x, q) = v(x, q)
\end{align*}$$
Let the subset $W_k(q)$ of $W$ be defined as the set of all possible realizations of $Q_k$ given $Q_0 = q$. Note that $W_k(q)$ is a finite set. Hence to compute the approximation $(U^n b)(x,q)$ of $v(x,q)$ we have to solve a dynamic programming problem with $n$ stages and the number of states at level $k$ equals $\#(X) \times \#(W_k(q))$. However to guarantee that the approximation is good, the number $n$ has to be very large, if the so-called scrap function $b$ on $X \times W$ is constant in the second argument (cf. [Martin (1967)], [Van Hee (1978), p. 137]).

If the horizon $n$ is large, then the number of elements in $W_n(q)$ is also very large and it turns out that only very few problems can be solved in this way.

To overcome these problems we introduce scrap functions $b$ on $X \times W$ (which are non-constant in the second argument and) which do not require such a large horizon. Further we introduce a parameter structure, such that the number of elements in $W_n(q)$ is relatively small, and we consider for this situation a relatively fast approximation method to compute for each $\varepsilon > 0$ a horizon $n$ such that $| (U^n b)(x,q) - v(x,q) | < \varepsilon$.

Finally we consider easy-to-handle strategies that do not require the knowledge of the value function (cf. 1.10) and that behave good in special situations.

2. Parameter structure. First we sketch the parameter structure and afterwards we consider an example satisfying this structure.

Let the set $X$ be a product space:

2.1. $X = \tilde{X} \times Y$

and let $R$ be a transition probability from $\tilde{X} \times A \times Y$ to $\tilde{X}$. Further we consider a transition probability $p$ from $\tilde{X} \times A \times \theta$ to $Y$ and we assume the following structure

2.2. $P(x',y' \mid x,y,a,\theta) = R(x' \mid x,a,y')p(y' \mid x,a,\theta)$

$(x,x' \in \tilde{X}, y,y' \in Y, a \in A \text{ and } \theta \in \Theta)$.

It is easy to verify that (provided that the denominator does not vanish)

$$T(x,y,a,(x',y'))(\theta) = \sum_{\theta'} \frac{P(y' \mid x,a,\theta)q(\theta)}{P(y' \mid x,a,\theta')q(\theta')}, \quad \theta \in \Theta.$$ 

In this section we shall assume that $\{L_1,L_2\}$ is a partition of $\tilde{X}$ and that

2.3. $p(y' \mid x,a,\theta) = p_1(y' \mid \theta)1_{L_1}(x) + p_2(y' \mid \theta)1_{L_2}(x)$,

(where $1_B(x) = 1$ if $x \in B$, = 0 otherwise).

Hence if $x \in L_1$ then the transition depends on a distribution with an unknown parameter and if $x \in L_2$ the transition distribution is completely known.
In this case we have for \( x \in L_1 \):

\[
2.4. \ i) \quad T(x,y),a,(x',y')(q)(\theta) = \frac{p_1(y'|\theta)q(u)}{\sum_{\theta} p_1(y'|\theta)q(\theta)}, \ \theta \in \Theta
\]

and for \( x \in L_2 \):

\[
2.4. \ ii) \quad T(x,y),a,(x',y')(q)(\theta) = q(\theta), \ \theta \in \Theta
\]

Hence the posterior distribution does not depend on the chosen actions, which reduces the number of elements in \( W_1(q) \), \( n \in \mathbb{N} \). From now on we assume that \( r(x,y,a) \) does not depend on \( y \) and we omit \( y \) in the notations. We continue with a motivation for this parameter structure and afterwards we consider an example.

The state definition of a system, in case the transition law is completely known, is not always appropriate if the transition law is incompletely known. For example in an inventory control model without backlogging the inventory level may be chosen as the state variable if the demand distribution is known. However if the demand distribution is unknown, then the sequence of successive inventory levels does not determine the sequence of successive demands and therefore we have to consider the demand in each period as a supplementary state variable.

So we consider the space \( \tilde{X} \) as the state space of the original model when the transition law is known and \( Y \) as the space of the supplementary state variables.

**Example.** 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 number of elements in the group is unknown. The service distribution is exponentially with a known and controllable parameter \( a \). Let \( y' \) be the number of arrivals in some period and let \( x \) be the queue length at the beginning of the period and \( x' \) at the end. Then, if \( c \cdot x + y' - x' \geq 0 \) 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 \( p(y'|x,a,\theta) \) does not depend on \( x \) and \( a \), and is the probability of a group of size \( y' \).

In case 2.3 holds, and also in more general situations, it can be proved that the model 1.1 is equivalent to a MDP specified by

\[
2.5. \quad ((\tilde{X} \times W),A,\tilde{P},\tilde{r})
\]

Hence the supplementary state variable, which was required only to save all information concerning the unknown parameter, disappears when we consider the posterior distribution as a state component.

Finally we note that the parameter structure given in 2.2 includes the original model. To see this, let \( Y := \tilde{X} \) and let...
3. Approximations. In this section we restrict us to the case where \( X, Y, A \) and \( \Theta \) are finite sets except for the last part where \( \Theta \) is an arbitrary set. We start with the study of upper and lower bounds on the value function (cf. 1.10). Afterwards we consider successive approximations and we discuss computational procedures to approximate the value function for a fixed prior distribution \( q \in W \). Finally we consider the case where \( \Theta \) is an arbitrary set and where we approximate the prior distribution \( q \) by another one \( \phi \) where \( \phi \) is concentrated on finitely many points. We provide bounds for the difference between the values \( v(x,q) \) and \( v(x,\phi) \). We start with some notations.

3.1. \( F := \{ f : X \to A \} \).

Hence each \( f \in F \) represents a stationary policy for the model with known transition 1. We identify each \( f \in F \) with the strategy \( \pi \in \Pi \) that corresponds to \( f \) in the following way:

\[
\pi_n(f(x_i) \mid x_0, y_0, a_0, \ldots, x_n, y_n) = 1 \quad \text{for all } x_i \in \tilde{X}, y_i \in Y, a_i \in A.
\]

Further we consider the subset \( \tilde{F} \subset F \) defined by

3.2. \( \tilde{F} := \{ f \in F \mid v(x,\theta,f) = v(x,\theta) \text{ for all } x \in \tilde{X} \text{ and some } \theta \in \Theta \} \).

Hence \( \tilde{F} \) contains all stationary policies that are optimal for some parameter \( \theta \in \Theta \).

Finally we define two functions on \( X \times W \):

3.3. i) \( w(x,q) := \sum_{\theta} v(x,\theta)q(\theta) \)

3.4. ii) \( l(x,q) := \max_{f \in F} \sum_{\theta} v(x,\theta,f)q(\theta) \)

(note that \( \sum_{\theta} v(x,\theta,f)q(\theta) = v(x,q,f) \)).

Theorem 3.1. The following properties hold

3.4. i) \( (U_n^w)(x,q) \leq v(x,q) \leq (U_n^l)(x,q) \) for \( n \in \mathbb{N} \).

ii) \( (U_n^l)(x,q) \) is nondecreasing and \( (U_n^w)(x,q) \) is nonincreasing in \( n \).

both with limit \( v(x,q) \) (note that \( U_0^l := b \)).

It turns out that in a lot of problems the bounds \( w \) and \( l \) are rather tight. However properties 3.4 i) and ii) give us the possibility to approximate \( v(x,q) \) for fixed \( q \in W \) as accurate as you like. Namely if we allow an error \( \varepsilon > 0 \) then we have to fix a horizon \( n \) and we compute \( (U_n^w)(x,q) - (U_n^l)(x,q) \). If this difference is too large we have to repeat the whole procedure for a larger horizon. Unfortunately the values we have to compute \( (U_n^w)(x,q) - (U_n^l)(x,q) \) are of no help for obtaining
(U_m^w)(x,q) - (U_m^x)(x,q) for m > n. So it would be nice to have a method to determine a good horizon in advance. In case we are dealing with the structure given in 2.3, we have such a method. To this end we introduce a new optimal reward operator, which is based on a stopping time σ:

3.5. \( \sigma(\omega) := \inf\{n > 0 \mid X_n(\omega) \in L_1\}, \quad \omega \in \Omega \).

The optimal reward operator \( U_\sigma \) is defined for bounded functions \( b: \tilde{X} \times W \to \mathbb{R} \) (\( b \) has to be measurable on \( W \)):

3.6. \( (U_\sigma b)(x,q) := \sup_{\pi \in \Pi} \mathbb{E}^\pi \left[ \sum_{n=0}^{\sigma-1} \beta^n x(X_n, A_n) + \beta^\sigma b(X_\sigma, Q_\sigma) \right] \)

(note that \( U_\sigma = U \) if \( L_1 = \tilde{X} \)).

**Theorem 3.2.** Let \( b(x,q) := \frac{1}{2}(w(x,q) + l(x,q)) \) (cf. 3.3).

i) Then

\[
|v(x,q) - (U^n b)(x,q)| \leq \beta^n S(q,n) \quad \text{if} \ x \in L_1 \\
\leq \beta^{n-1} S(q,n-1) \quad \text{if} \ x \in L_2
\]

where

3.7. \( S(q,n) := \frac{1}{2} \min_{y_1,\ldots,y_n \in Y} \min_{f \in F} \max_{\theta \in \bar{\Theta}} \{v(x,\theta) - v(x,\theta,f)\} \).

ii) The sequence \( \{S(q,n), n = 0,1,2,\ldots\} \) is nonincreasing and \( \lim_{n \to \infty} S(q,n) = 0 \). This convergence is exponentially fast.

(For a proof of these statements cf. [van Hee (1978), th. 6.5 and th. 7.2]).

To use theorem 3.2 to obtain a horizon estimation we have to compute the values \( v(x,\theta) \) and \( v(x,\theta,f) \) for \( x \in \tilde{X}, \theta \in \Theta, f \in \bar{F} \) in advance. It turns out that, if the functions \( \theta \to v(x,\theta,f), x \in \tilde{X}, f \in \bar{F} \) are smooth, then these computations are rather quick. The computation of \( (U_\sigma b)(x,q) \) for \( x \in L_2 \) can be done by solving an ordinary dynamic programming problem with all states of \( L_1 \) absorbing. Since if \( X_0 \in L_2 \) we have \( X_n \in L_2 \) for \( n < \sigma \) and \( Q_\sigma = Q_0 = q \). Hence

3.8. \( (U_\sigma b)(x,q) = \sup_{\pi \in \Pi} \mathbb{E}^\pi \left[ \sum_{n=0}^{\sigma-1} \beta^n x(X_n, A_n) + \beta^\sigma b(X_\sigma, Q_\sigma) \right] \)

(note that the expectation does not depend on \( q \) since for all states \( X_n \in L_2 \) the transition law is known).
We conclude this section with a theorem concerning an error estimate for discretizing the prior distribution. Let \( \Theta \) be an arbitrary (measurable) set and let \( q \) be a probability on \( \Theta \). Further let \( \{B_1, B_2, \ldots, B_n\} \) be a partition of \( \Theta \) and let \( b_j \in B_j, j = 1, 2, \ldots, n \). Then we define another probability \( \phi \) on \( \Theta \) by

\[
\phi([b_j]) := q(B_j), \quad j = 1, 2, \ldots, n.
\]

So \( \phi \) is discretization of \( q \). For computational reasons it is nice to have a finite parameter space or equivalently a prior distribution concentrated on finitely many points. However in practice it is unrealistic to consider these prior distributions. The following theorem gives us the opportunity to discretize the prior distribution such that the approximation is as good as we like.

**Theorem 3.3.** Let \( q \in W \) and \( \phi \in W \) be defined by 3.10. Then

\[
\max_{x \in X} |v(x, q) - v(x, \phi)| \leq \frac{\text{span}(r)}{1 - \beta} \sum_{j=1}^{n} \int_{B_j} \frac{\Delta(\theta, b_j)}{1 - \beta + \beta \Delta(\theta, b_j)}
\]

where

\[3.11\]

\[3.11\]

\[3.11\]

\[\Delta(\theta, \hat{\theta}) := \sum_{y} |p_1(y|\theta) - p_1(y|\hat{\theta})|,\]

\[\text{span}(r) := \max_{x \in X} r(x, a) - \min_{x \in X} r(x, a) \quad \text{for } a \in A.
\]

This theorem gives us also an upper bound for \(|v(x, q) - v(x, \hat{\theta})|\) where \( \hat{\theta} := \int \theta q(d\theta) \), in case \( \theta \) is an interval on \( \mathbb{R} \). Here \( \hat{\theta} \) is the prior Bayes estimate of the parameter. Namely,

\[
\max_{x \in X} |v(x, q) - v(x, \hat{\theta})| \leq \frac{\text{span}(r)}{1 - \beta} \int_{\Theta} \frac{\Delta(\theta, \hat{\theta})}{1 - \beta + \beta \Delta(\theta, \hat{\theta})}.
\]

Finally we note that any strategy that chooses at each state \((x, q) \in \tilde{X} \times W\) an action \( a^* \) that maximizes the function

\[
a \rightarrow \{r(x, a) + \beta \sum_{x'} P(x'|x, a, \theta)q(\theta)v(x', T_{x, a, x'}(q))\}
\]

on the set \( A \), is optimal. Hence if we compute \( v(x', T_{x, a, x'}(q)) \) for all \( a \in A, x' \in \tilde{X} \) then we can determine an optimal action, in \((x, q)\). This is a very time consuming procedure. Therefore we are looking for easy-to-handle strategies which behave good.
gin (1978)] the following strategy is considered for the average return case:

"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 action corresponding to this strategy in the actual state.
Repeat this procedure at the next stage".

In the discounted total return case an optimal action is found as a maximizer of the
function

$$a \rightarrow r(x,a) + \beta \sum_{x'} P(x' \mid x,a,\theta) v(x',\theta) =: F(x,\theta,a)$$

and in the average return case a similar function $F$ has to be maximized (in case $\tilde{X}$ and
$A$ are finite). So the above mentioned authors are maximizing at each stage $a \rightarrow P(x,\hat{\theta},a)$
where $\hat{\theta}$ is the estimation. If $\theta$ is an interval on the real line the Bayes estimate of
$\theta$ in state $(x,q) \in \tilde{X} \times W$ would be

$$\hat{\theta} = \int \theta q(d\theta).$$

In the average return case this strategy is optimal under some conditions guaranteeing
that the estimators are consistent.

We suggest another heuristic to obtain a strategy:

"At each stage, in state $(x,q) \in \tilde{X} \times W$, compute a maximizer of the function

$$a \rightarrow \sum_{\theta} q(\theta) F(x,\theta,a).$$

Where the function $F$ must have the property that a maximizer of $a \rightarrow F(x,\theta,a)$ gives an
optimal action in case $\theta$ is the true parameter value (so for example the function $F$
de fined in 4.1 produces such a strategy).

We call this heuristic strategy a Bayesian equivalent rule (BER). In the average re­
turn case these strategies are optimal under some conditions guaranteeing that $Q_n$
con­
vrges to a degenerate distribution. We consider some models where a BER is optimal or
where it behaves good. Finally we consider a bound for $v(x,q) - v(x,q,\pi^*)$ where $\pi^*$ is
a the BER defined by the function $F$ given in 4.1.

Example 4.1. Linear system with quadratic cost and independent disturbances with un­
known distribution (we consider here Euclidean spaces instead of countable sets). Let
$\tilde{X} = Y = \mathbb{R}^n$, $A = \mathbb{R}^m$. Let $C$ be a $n \times n$-matrix, $B$ a $n \times m$-matrix, $D$ a nonnegative de­
finite $n \times n$-matrix and $G$ a positive definite $m \times m$-matrix. The transition law is given
by

$$R((Cx + Ba + y') \mid x,a,y') := 1 \quad \text{for all } x \in \tilde{X}, a \in A, y' \in Y.$$

and $p(y' \mid x,a,\theta) := p_1(y' \mid \theta)$ is a probability density with respect to the Lebesgue mea­
ure on $Y$ (cf. 2.2). The reward function is given by
The only assumption we need is that the function

\[ \theta \mapsto \int |y_1 y_2| p_1(y|\theta) dy \]

is bounded on \( \Theta \).

In this model the BER given in 4.1 is optimal.

**Example 4.2. Inventory control model with backlogging and without fixed set up costs.**

\( X_n \) is the inventory level before ordering, \( A_n \) the inventory level after ordering and \( Y_n \) is the demand during the period \( n \). Let \( \tilde{X} = Y = \mathbb{R} \). Here the admissible actions depend on the state: \( A(x) = [x, \infty) \). It means that if the inventory level is \( x \) then the decisionmaker may change the inventory level by ordering only (and not by disposing inventory).

The transition law is given by:

\[ R(\{a - y'\} | x, a, y') = 1 \text{ for all } x \in \tilde{X}, a \in A(x), y' \in Y \]

and \( p(y'|x,a,s) = p_1(y'|s) \) is a density with respect to the Lebesgue measure on \( \mathbb{R} \).

The reward function is

\[ r(x,a) = - (hx^+ + px^- - c(a - x)) \]

where \( h \) is the holding cost, \( p \) the penalty cost (for being out of stock) and \( c \) the production cost (per unit) (\( x^+ := \max(0,x), x^- := -\min(0,x) \)).

Let \( v(x,q,\pi^*) \) be the discounted total return in case \( \pi^* \) is the BER. Then the following inequality holds:

\[ v(x,q) - v(x,q,\pi^*) \leq \left( \frac{1}{1 - \beta} h + k \right) \{(x - s(q))^+ + \sum_{n=1}^{\infty} \beta^n q_n [(s(Q_{n-1}) - s(Q_n) - Y_n)^+] \} \]

where

\[ s(q) := \inf\{a \in \mathbb{R} | \int_0^a \left\{ \int p_1(y|\theta) q(d\theta) \right\} dy \geq \frac{p - \frac{1 - \beta}{\beta} k}{p + h} \} \]

Let \( \delta := \sup s(\theta) - \inf s(\theta) \) (here \( s(\theta) \) is defined by 4.3, where \( q \) is degenerate at \( \theta \)).

Then we obtain, using 4.2, the following more appealing inequality:

\[ v(x,q) - v(x,q,\pi^*) \leq \left( \frac{1}{1 - \beta} h + k \right) \{(x - s(q))^+ + \frac{\beta}{1 - \beta} \int_0^\delta \left\{ \int p_1(y|\theta) q(d\theta) \right\} dy \} \]

Therefore, if \( \int_0^\delta p_1(y|\theta) dy = 0 \) for all \( \theta \in \Theta \), then the BER is optimal.

Sometimes it is possible to compute the upperbound in 4.2 exactly.

The BER defined by 4.1 has the following property.
Theorem 4.1. Let $\pi^*$ be the strategy defined by the BER of the form 4.1. Then

$$v(x,q) - v(x,q,\pi^*) \leq \frac{1}{1 - \beta} \min_{f \in F} \sum_{\theta} q(\theta) \psi(\theta,f)$$

where

$$\psi(\theta,f) := \max_{x} \{v(x,\theta) - r(x,f(x)) - \beta \sum_{x'} P(x'|x,f(x),\theta) v(x',\theta)\}.$$ 

However, if under $\pi^*$, it is guaranteed that $Q_n$ converges to a degenerate distribution then

$$\lim_{n \to \infty} \mathbb{E}_{x,q}^\pi [v(X_n, Q_n) - v(x_n, Q_n, \pi^*)] = 0.$$ 

Finally we remark that the BER considered in 4.1 behaves very good in a lot of numerical examples.

References.


van Hee, K.M., Bayesian Control of Markov chains. Amsterdam Mathematical Centre Tracts 95 (1978).


