Approximations in Bayesian controlled Markov chains

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 76-15

Approximations in Bayesian Controlled Markov Chains

by

K.M. van Hee

Eindhoven, October 1976

The Netherlands

Approximations in Bayesian Controlled Markov Chains

by

K.M. van Hee

0. Summary

A class of Markov decision processes is considered with a finite state and action space and with an incompletely known transition mechanism. The controller is looking for a strategy maximizing the Bayesian expected total discounted return. In section 2 approximations are given for this value and in section 3 we indicate how to compute the value for a fixed prior distribution.

1. Introduction and Preliminaries

For a detailed description of the model we refer to van Hee (1976A), here we only give a sketch. For statements without proof see also van Hee (1976A). Consider a Markov decision process with a finite *state space* S and a finite *action space* A. Let $r: S \times A \to \mathbb{R}$ be the *reward function*.
Let X_n be the state of the system at time n. There is a subset $B \subset S$ such that if $X_n \in B$ the next state is partially determined by the outcome of a random variable Y_{n+1}, where $\{Y_n, n = 1,2,3,\ldots\}$ is a sequence of i.i.d. random variables not controllable by the decisionmaker. The process $\{Y_n, n = 1,2,3,\ldots\}$ is called the *external process* and has a finite state space E. If and only if $X_n \in B$ then Y_{n+1} becomes *visible* to the decisionmaker. Let P be a transition probability from $S \times A \times E$ to S such that

$$\mathbb{P}[X_{n+1} = t \mid X_n = s,\ A_n = a,\ Y_{n+1} = y] = P(t|s,a,y) ,$$

where A_n is the action at time n. (For $s \in S\backslash B$ $P(t|s,a,\cdot)$ is constant and we omit the dependence on y in this case). Only the distribution of the external process; i.e. $p(y|\theta) := \mathbb{P}_\theta[Y_{n+1} = y]$, depends on an unknown parameter $\theta \in \Theta$ where Θ is a finite *parameter* space. Note that for each fixed $\theta \in \Theta$ the process forms an ordinary Markov decision process with transition probability:

$$\mathbb{P}_\theta[X_{n+1} = t \mid X_n = s,\ A_n = a] = \sum_{y \in E} P(t|s,a,y).p(y|\theta) .$$

Examples of such a model can be found in inventory control, where Y_n is the demand in period $(n-1,n]$ and also in queueing models, where Y_n is the number of newcomers.

Let Π be the set of all strategies based on the visible histories (i.e. for each $\pi \in \Pi$ the action A_n may depend on $X_0,\ldots,X_n$, $A_0,\ldots,A_{n-1}$ and on Y_k if $X_{k-1} \in B$, $k = 1,2,\ldots,n$).
For each starting state $s \in S$, each $\pi \in \Pi$ and $\theta \in \Theta$ we have a random process $\{(X_n,A_n,Y_{n+1}),\ n = 1,2,3,\ldots\}$ and a probability $\mathbb{P}^{\pi}_{s,\theta}[\ ]$ on the sample space. (The expectation w.r.t. this probability is denoted by $\mathbb{E}^{\pi}_{s,\theta}$.)
The *Bayesian expected discounted total return* $v(s,q,\pi)$ w.r.t. an *prior distribution* q on θ is defined by

$$v(s,q,\pi) := \sum_{\theta} \mathbb{E}^{\pi}_{s,\theta}\Big[\sum_{n=0}^{\infty} \beta^n r(X_n,A_n)\Big].q(\theta),\quad s \in S,\ \pi \in \Pi\ ,$$

where $\beta \in [0,1)$ is the discount factor.
The set of all distributions on θ is denoted by W, and the function $v\colon S \times W \to \mathbb{R}$, defined by $v(s,q) := \sup_{\pi\in\Pi} v(s,q,\pi)$, is called the *value function*.
We define a sequence of stopping times:

$$\sigma_1 := \inf\{n \geq 0;\ X_n \in B\}$$

$$\sigma_k := \inf\{n > \sigma_{k-1};\ X_n \in B\},\ k = 2,3,4,\ldots$$

$$\tau_k := \sigma_k + 1\ ,\ k = 1,2,3,\ldots\ .$$

The Bayes criterion allows us to consider the parameter $\theta \in \Theta$ as a random variable Z with distribution q on θ.
Given $q \in W$, $s \in S$ and $\pi \in S$ we have a probability $\mathbb{P}^{\pi}_{s,q}$ on the sample space of the process $\{Z,(X_n,A_n,Y_{n+1}),\ n = 0,1,2,\ldots\}$ and for each event C defined by

$$C := \{X_0 = s_0,A_0 = a_0,Y_1 = y_1,\ldots,X_n = s_n,A_n = a_n,Y_{n+1} = y_{n+1}\}$$

we have

$$\mathbb{P}^{\pi}_{s,q}[C] = \sum_{\theta\in\Theta} \mathbb{P}^{\pi}_{s,\theta}[C]q(\theta)$$

(the expectation w.r.t. $\mathbb{P}^{\pi}_{s,q}$ is denoted by $\mathbb{E}^{\pi}_{s,q}$).
We define on the event $\{\tau_n < \infty\}$ for $s \in S$, $q \in W$, $\pi \in \Pi$:

$$\phi_n(\theta) := \mathbb{P}^{\pi}_{s,q}[Z = \theta \mid Y_{\tau_1},\ldots,Y_{\tau_n}]$$

and

$$Q_n(\theta) := \phi_k(\theta) \text{ on } \{\tau_k \leq n < \tau_{k+1}\}\ .$$

The vector valued process $\{Q_n\}$ with $Q_n := \{Q_n(\theta),\ \theta \in \Theta\}$ is called the *Bayes process*.

Note that, since the values of the external process are not influenced by the starting state s and the strategy π, we have that $\phi_n(\theta)$ does not depend on s and π on $\{\tau_n < \infty\}$, and evenso for $Q_n(\theta)$ if $B = S$.

If we are in the situation that expectations or conditional expectations do not depend on s and π we omit these sub and superscript.

We sometimes need the following conditions:

(A) for all $s \in S$, $\pi \in \Pi$ and $\theta \in \Theta$

$$\mathbf{P}^{\pi}_{s,\theta}[\bigcap_{n=1}^{\infty} \{\tau_n < \infty\}] = 1 .$$

(Note that $B = S$ implies (A).)

(B) For each pair $\theta, \hat{\theta} \in \Theta$ there is a $y \in E$ such that

$$p(y|\theta) \neq p(y|\hat{\theta}) .$$

(The only place where (B) is used is in the proof of the following theorem.)

Theorem 1. Assume (A,B). Thenfor all $s \in S$, $q \in W$ and $\pi \in \Pi$ it holds that

$$\lim_{n\to\infty} Q_n(\theta) = \delta_{z,\theta} \quad \mathbf{P}_{s,q}\text{-a.s.}$$

We need some notations

$p: E \times W \to [0,1]$ such that

$$p(y,q) := \sum_{\theta\in\Theta} q(\theta).p(y|\theta) ,$$

$T: W \times E \to W$ such that

$$T_y(q)(\theta) := \frac{p(y|\theta)q(\theta)}{p(y,q)} \text{ if } p(y,q) > 0, := q(\theta) \text{ otherwise} .$$

We may reduce our Bayesian decision problem to a *discounted dynamic program* with state space $S \times W$, action space A and reward function r, as stated in theorem 2.

Theorem 2. The value function v is the unique solution of the functional equation

$$v(s,q) = \max_{a\in A} \{r(s,a) + \beta \sum_{y\in E} \sum_{t\in S} P(t|s,a,y)p(y,q)v(t,T_y q)\}, \quad s \in B$$

$$= \max_{a\in A} \{r(s,a) + \beta \sum_{t\in S} P(t|s,a)v(t,q)\}, \quad s \in S\backslash B \ .$$

Corollary 1. There is an optimal strategy π^* which is *stationary*, i.e. there is a function $g: S \times W \to A$ such that π^* chooses action $g(s,q)$ in $(s,q) \in S \times W$.

2. Approximations

In this section we shall give some approximations for $v(s,q)$, $s \in S$ and a fixed prior $q \in W$. In section 3 we consider the computational aspects. We identify each $\theta \in \Theta$ with the degenerated distribution at θ. Hence $v(s,\theta)$ is the optimal value of the Markov decision process if s is the starting state and θ is known. Let

2.1. $$M := \{f \mid f: S \to A\}$$

be the set of *Markov policies* and identify the strategy $\pi \in \Pi$ that chooses action $f(s)$, $f \in M$ in state (s,q) with f.
Further let

$$F_\theta := \{f \in M \mid v(s,\theta) = v(s,\theta,f) \text{ for all } s \in S\}, \quad \theta \in \Theta$$

and $c: \Theta \to M$ be such that $c(\theta) \in F_\theta$, $\theta \in \Theta$. We define

2.2. i) $$F := \bigcup_{\theta\in\Theta} F_\theta$$

ii) $$\bar{F} := \{f \in M \mid f = c(\theta) \text{ for some } \theta \in \Theta\} \ .$$

On $S \times W$ we define the following functions:

2.3. i) $$w(s,q) := \sum_{\theta\in\Theta} v(s,\theta)q(\theta)$$

ii) $$\ell(s,q) := \max_{f\in F} \sum_{\theta\in\Theta} v(s,\theta,f)q(\theta)$$

iii) $$\bar{\ell}(s,q) := \max_{f\in\bar{F}} \sum_{\theta\in\Theta} v(s,\theta,f)q(\theta) \ .$$

Lemma 3.

i) $\bar{\ell}(s,q) \le \ell(s,q) \le v(s,q) \le w(s,q)$ for all $s \in S$, $q \in W$.

ii) Let (A,B) hold, then for all $t \in S$, $\pi \in \Pi$ and $q \in W$

$$\lim_{n\to\infty} \max_{s\in S} \{w(s,Q_n) - \bar{\ell}(s,Q_n)\} = 0, \ \mathbb{P}^{\pi}_{t,q}\text{-a.s.}$$

Proof.

i) $$\bar{\ell}(s,q) \le \ell(s,q) = \max_{f\in F} v(s,q,f) \le \sum_{\theta\in\Theta} q(\theta) \sup_{\pi\in\Pi} v(s,\theta,\pi) = w(s,q) .$$

ii) By theorem 1 we have

$$\lim_{n\to\infty} w(s,Q_n) = \lim_{n\to\infty} \sum_{\theta\in\Theta} v(s,\theta)Q_n(\theta) = v(s,Z), \ \mathbb{P}^{\pi}_{t,q}\text{-a.s.} .$$

Note that

$$\left|\max_{f\in F} \sum_{\theta} Q_n(\theta)v(s,\theta,f) - \max_{f\in F} v(s,Z,f)\right| \le \max_{f\in F} \left|\sum_{\theta} Q_n(\theta)v(s,\theta,f) - v(s,Z,f)\right| .$$

Hence

$$\lim_{n\to\infty} \bar{\ell}(s,Q_n) = v(s,Z), \ \mathbb{P}^{\pi}_{t,q}\text{-a.s.} \qquad \square$$

Define two functions:

2.4. i) $$\varphi(s,a,\theta) := r(s,a) + \beta \sum_{t\in S} \sum_{y\in E} P(t|s,a,y)p(y|\theta)v(t,\theta) - v(s,\theta), \quad s \in S, \ a \in A, \ \theta \in \Theta .$$

ii) $$\varphi(s,q) := \max_{a\in A} \sum_{\theta\in\Theta} \varphi(s,a,\theta)q(\theta), \quad s \in S, \ q \in W .$$

Note that $\varphi(s,a,\theta) \le 0$ for all $s \in S$, $a \in A$ and $\theta \in \Theta$ and note also that $\varphi(s,q) = 0$ if q is a degenerated distribution.

Lemma 4.

i) $$v(s,q) \ge w(s,q) + \frac{1}{1-\beta} \max_{a\in A} \sum_{\theta\in\Theta} \min_{x\in S} \varphi(x,a,\theta)q(\theta) .$$

ii) $$v(s,q) \ge w(s,q) + \frac{1}{1-\beta} \max_{f\in F} \sum_{\theta\in\Theta} \min_{x\in S} \varphi(x,f(x),\theta)q(\theta) .$$

iii) if $B = S$ then

$$\operatorname*{span}_{s} \{w(s,q) - v(s,q)\} \le \mathbf{E}_q\Big[\sum_{n=0}^{\infty} \beta^n \operatorname*{span}_{s} \varphi(s,Q_n)\Big] . \ ^{*}$$

* $\operatorname*{span}_{x} f(x) := \sup_{x} f(x) - \inf_{x} f(x)$.

Proof. By 2.4 we have for $a \in A$:

$$\sum_{\theta\in\Theta} q(\theta)v(s,\theta) + \varphi(s,q) \geq r(s,a) + \beta \sum_{\theta}\sum_{t}\sum_{y} P(t|s,a,y)p(y|\theta)q(\theta)v(t,\theta).$$

Note that $p(y|\theta)q(\theta) = (T_y q)(\theta)p(y,q)$. Hence by substituting 2.3i) we have

$$w(s,q) + \varphi(s,q) \geq r(s,a) + \beta \sum_{t}\sum_{y} P(t|s,a,y)p(y,q)w(t,T_y q), \quad \text{if } s \in B$$

$$\geq r(s,a) + \beta \sum_{t} P(t|s,a)w(t,q) \quad , \quad \text{if } s \in S\backslash B.$$

Let π be a stationary strategy (see corollary 1.1) and define

$$f(s,q) := \mathbf{E}^{\pi}_{s,q}[w(X_1,Q_1)], \; s,q \in S \times W$$

then

$$r(s,q) \leq \varphi(s,q) + w(s,q) - \beta f(s,q) \; .$$

By the Markov property we have for all $s,q \in S \times W$:

$$f(X_n,Q_n) = \mathbf{E}_{s,q}[w(X_{n+1},Q_{n+1}) \mid X_n,Q_n], \; \mathbf{P}^{\pi}_{s,q}\text{-a.s.}$$

Hence

2.5. $$v(s,q,\pi) \leq \mathbf{E}^{\pi}_{s,q}[\sum_{n=0}^{\infty} \beta^n \varphi(X_n,Q_n)] + \mathbf{E}^{\pi}_{s,q}[\sum_{n=0}^{\infty} \beta^n w(X_n,Q_n)] +$$

$$- \mathbf{E}^{\pi}_{s,q}[\beta \sum_{n=0}^{\infty} \beta^n w(X_{n+1},Q_{n+1})] = w(s,q) + \mathbf{E}_{s,q}[\sum_{n=0}^{\infty} \beta^n \varphi(X_n,Q_n)].$$

Let $\tilde{\pi}$ be the strategy that chooses in $(s,q) \in S \times W$ a fixed action a, maximizing $\sum_{\theta} q(\theta)\varphi(s,a,\theta)$. Note that $\tilde{\pi}$ is stationary and note also that equality holds in 2.4 if $\pi = \tilde{\pi}$.

We first prove iii).

Let π^* be a stationary optimal strategy, then

2.6. $$v(s,q) = v(s,q,\pi^*) \leq w(s,q) + \mathbf{E}^{\pi^*}_{s,q}[\sum_{n=0}^{\infty} \beta^n \max_{x\in S} \varphi(x,Q_n)] \; .$$

But

2.7. $$v(s,q) \geq v(s,q,\tilde{\pi}) \geq w(s,q) + \mathbf{E}^{\tilde{\pi}}_{s,q}[\sum_{n=0}^{\infty} \beta^n \min_{x\in S} \varphi(x,Q_n)] \; .$$

Remark that under the condition B = S the distribution of Q_n is independent of $s \in S$ and $\pi \in \Pi$, hence iii) is a direct consequence of 2.6 and 2.7.
To prove i) and ii) note that

$$\min_{x\in S} \varphi(x,q) = \min_{x\in S} \max_{f\in F} \sum_{\theta} q(\theta)\varphi(x,f(x),\theta) = \max_{f\in F} \min_{x\in S} \sum_{\theta} q(\theta)\varphi(x,f(x),\theta) \geq$$

$$\geq \max_{f\in F} \sum_{\theta} q(\theta) \min_{x\in S} \varphi(x,f(x),\theta) \geq \max_{a\in A} \sum_{\theta} q(\theta) \min_{x\in S} \varphi(x,a,\theta) .$$

Further note that the last two expressions are convex functions on W so by Jensen's inequality applied to the right hand side of 2.7 we have the desired result. □

Remark 2.8. By the proof of lemma 4 we see that the lowerbound given in ii) is greater than or equal to the lowerbound of i), but it requires more work to compute it. Further note that, if (A,B) holds

$$\lim_{n\to\infty} \max_{f\in F} \sum_{\theta} \min_{x\in S} \varphi(x,f(x),\theta) Q_n(\theta) = 0, \ \mathbb{P}^{\pi}_{s,q}\text{-a.s.}$$

since

$$Q_n(\theta) \to \delta_{Z,\theta}, \ \mathbb{P}^{\pi}_{s,q}\text{-a.s.}$$

We introduce now an operator U working on the space G of bounded measurable functions on S × W (measurable w.r.t. the Borel σ-field on S × W):
Let $f \in G$:

$$2.9. \qquad (Uf)(s,q) := \sup_{\pi\in\Pi} \mathbf{E}^{\pi}_{s,q}\Big[\sum_{n=0}^{\tau_1-1} \beta^n r(X_n,A_n) + \beta^{\tau_1} f(X_{\tau_1},Q_{\tau_1})\Big] .$$

Note that for $f \in G$ Uf is continuous on W since

$$|(Uf)(s,q) - (Uf)(s,\varphi)| \leq \{\frac{M}{1-\beta} + M'\} \sum_{\theta} |q(\theta) - \varphi(\theta)|$$

for $q,\varphi \in W$, $|r(s,a)| \leq M$, $|f| \leq M'$.
Note further that G is a Banach space w.r.t. the supremum norm.
In Wessels (1974) and van Nunen (1976) a class of operators of this type is studied for models with a finite respectively countable state space. They both prove the following theorem. For our situation it is proved in van Hee (1975) in a similar way.

Theorem 5. The operator U (defined in 2.9) is monotone and contracting. The value function v is the unique fixed point of U in G.

The next theorem is important for successive approximations. Let us assume that $\tilde{v}$ is an approximation of v and that the difference $|\tilde{v} - v|$ is bounded by a function ε.

Theorem 6. Let v be the value function and let $\tilde{v}$ and $\varepsilon \in G$, such that

$$|v(s,q) - \tilde{v}(s,q)| \leq \varepsilon(s,q) \quad \text{for all } s \in S,\ q \in W$$

then it holds that

$$|v(s,q) - (U^n\tilde{v})(s,q)| \leq \sup_{\pi\in\Pi} \mathbf{E}^{\pi}_{s,q}[\beta^{\tau_n}\varepsilon(X_{\tau_n},Q_{\tau_n})] .$$

Proof. First we define the operator L: G → G by

$$(Lf)(s,q) := \sup_{\pi\in\Pi} \mathbf{E}^{\pi}_{s,q}[\beta^{\tau_1} f(X_{\tau_1},Q_{\tau_1})],\ f \in G,\ s \in S,\ q \in W$$

(it is easy to verify that Lf is continuous on W, so $Lf \in G$). It holds that

$$(U(v+\varepsilon))(s,q) \leq (Uv)(s,q) + (L\varepsilon)(s,q) \leq v(s,q) + (L\varepsilon)(s,q)$$

and therefore

$$(U^n(v+\varepsilon))(s,q) \leq v(s,q) + (L^n\varepsilon)(s,q)$$

and in the same way

$$(U^n(v-\varepsilon))(s,q) \geq v(s,q) - (L^n\varepsilon)(s,q) .$$

So, again by the monotonicity of U, we have

$$|(U^nv)(s,q) - v(s,q)| \leq (L^n\varepsilon)(s,q) .$$

To complete the proof we have to verify that

$$(L^n\varepsilon)(s,q) \leq \sup_{\pi\in\Pi} \mathbf{E}^{\pi}_{s,q}[\beta^{\tau_n}\varepsilon(X_{\tau_n},Q_{\tau_n})]$$

for the rather technical proof of this statement we refer to van Hee (1976A). □

Corollary 7. Suppose that B = S. Let $\tilde{v} \in G$ and let $\varepsilon: W \to \mathbb{R}$ be a bounded measurable function. If

$$|v(s,q) - \tilde{v}(s,q)| \leq \varepsilon(q)$$

then

$$|v(s,q) - (U^n\tilde{v})(s,q)| \leq \mathbf{E}_q[\beta^n \varepsilon(Q_n)] .$$

To prove this statement note that B = S implies $\tau_n = n$ and that the distribution of Q_n is independent of the starting state and the strategy.

Corollary 8. Suppose that B = S. Let $\tilde{v}(s,q) := \frac{1}{2}\{w(s,q) + \ell(s,q)\}$ and

$$\varepsilon(q) := \tfrac{1}{2} \min_{f\in F} \sum_{\theta\in\Theta} \max_{x\in S} \{v(x,\theta) - v(x,\theta,f(x))\}q(\theta) .$$

Then:

i) $$|v(s,q) - (U^n\tilde{v})(s,q)| \leq \mathbf{E}_q[\beta^n \varepsilon(Q_n)] ,$$

ii) $$\mathbf{E}_q[\varepsilon(Q_n)] \geq \mathbf{E}_q[\varepsilon(Q_{n+1})] ,$$

iii) $$\lim_{n\to\infty} \mathbf{E}_q[\varepsilon(Q_n)] = 0 .$$

Proof.

i) Note that

$$|v(s,q) - \tilde{v}(s,q)| \leq \tfrac{1}{2}\{w(s,q) - \ell(s,q)\} \leq \varepsilon(q) .$$

ii) Note that $\varepsilon(q)$ is a concave function on W. Since $\{Q_n, n \in \mathbb{N}\}$ forms a martingale (see van Hee (1976A)) we have that $\{\varepsilon(Q_n), n \in \mathbb{N}\}$ forms a supermartingale.

iii) By theorem 1 we have $\mathbb{P}_q$-a.s.

$$\lim_{n\to\infty} \varepsilon(Q_n) = \tfrac{1}{2} \min_{f\in F} \max_{x\in S} \{v(x,Z) - v(x,Z,f(x))\} = 0 .$$ □

Remark 2.10. Let B = S and define

i) $$\varepsilon(q) := \tfrac{1}{2} \frac{1}{1-\beta} \max_{f\in F} \sum_{\theta\in\Theta} \min_{x\in S} \varphi(x,f(x),\theta)q(\theta) ,$$

ii) $$\tilde{v}(s,q) := w(s,q) + \varepsilon(q) .$$

Then the three statements of corollary 8 hold also. The proof proceeds along the same lines, using lemma 4 and remark 2.8.
If in corollary 8 ℓ is replaced by $\bar{\ell}$ and F by $\bar{F}$ the statements i) and iii) remain true.

3. Computational aspects and additional remarks

The approximations given in section 2 are of interest for computations if we are prepared to determine the sets F and $\{v(s,f,\theta) \mid s \in S, \theta \in \Theta, f \in F\}$ (or F replaced by $\bar{F}$). Let $k := \#(\Theta)$ then the determination of F requires the solution of k ordinary Markov decision problems with a finite state and action space and the determination of all optimal policies. If $n := \#(F)$ (or $\#(\bar{F})$) then we have to solve $(k-1)n$ systems of linear equations to determine the second set.
If there is a $f \in M$ which is optimal for all $\theta \in \Theta$ then $v(s,q) = w(s,q)$ for all $s \in S$, $q \in W$. For separable value functions, i.e. for models with $v(s,\theta) = h(s) + g(\theta)$, it holds that $\operatorname{span}_s \varphi(s,q) = 0$, hence by lemma 4iii) we have that $\operatorname{span}_s\{v(s,q) - w(s,q)\} = 0$. In van Hee (1976B) a class of problems, including some inventory control models, is considered with this structure.

For each $q \in W$ we define

$$W_n(q) := \{\varphi \in W \mid \varphi = T_{y_n}(T_{y_{n-1}}(\ldots(T_{y_1}(q))\ldots), y_1,\ldots,y_n \in E\}$$

hence $W_n(q)$ is the set of all n-stage posterior distributions of q. The sets $W_n(q)$ and $W_m(q)$, $n \neq m$ are in general not disjoint (see van Hee (1976A)).
For a fixed $q \in W$ it follows from section 2, that, loosely speaking, the approximations of $v(s,\varphi)$ for $\varphi \in W_n(q)$ are better if n is large.
Since $(U^n\tilde{v})(s,q)$ requires only the values of $\tilde{v}(s,\varphi)$ for $\varphi \in W_n(q)$, $s \in S$ we may approximate $v(s,\varphi)$ by $\tilde{v}(s,q)$ on $W_n(q)$ and then by backward induction we can determine $(U^n\tilde{v})(s,q)$. The only problem is the determination of n, the horizon.
For models with B = S corollary 8i) shows that the error determination is rather easy: we have only to compute $\beta^n E_q[\varepsilon(Q_n)]$, which requires the determination of $\varepsilon(\varphi)$ for all $\varphi \in W_n(q)$, to check whether horizon n is sufficiently accurate or not. If B is a proper subset of S and if (A,B) holds a similar result holds since

$$\sup_{\pi\in\Pi} \mathbf{E}^{\pi}_{s,q}[\beta^{\tau_n}\varepsilon(Q_{\tau_n})] \leq \beta^n \mathbf{E}_q[\varepsilon(Q_{\tau_n})] ,$$

viz. the distribution of Q_{τ_n} depends only on q.

In van Hee (1976A) two algorithms are presented based on these arguments, for models with B = S and for models where B consists of only one state. Also numerical results are given there and attention is paid to the determination of optimal actions.

In Martin (1967) the usual method of successive approximations is proposed with a terminal function t: $S \to \mathbf{R}$. In our terminology Martin approximates $v(s,q)$ by $(U^n t)(s,q)$. The difficulty of this method is that the choice of the horizon must be made on the error estimate $\frac{1}{2}\beta^n \frac{\bar{M} - \underline{M}}{1 - \beta}$, where

$$\bar{M} := \max_{s,a} r(s,a), \quad \underline{M} := \min_{s,a} r(s,a) .$$

Satia and Lave (1973) also suggest the use of upper and lower bounds for $v(s,\varphi)$, $\varphi \in W_n(q)$. It is easy to see that their bounds are worse than ours (see van Hee (1976A)).

Literature

van Hee, K.M. (1976A); Bayesian control of Markov chains, to appear.

van Hee, K.M. (1976B); Adaptive control of special structured Markov chains, to appear.

Martin, J.J. (1967); Bayesian decision problems and Markov chains, Wiley, New York.

van Nunen, J.A.E.E. (1976); Contracting Markov decision processes, MC-tract, Amsterdam.

Satia, J.R. and R.E. Lave (1973); Markov decision processes with uncertain transition probabilities, Operations Research 21.

Wessels, J. (1974); Stopping times and Markov Programming. Proceedings of 1974 E.M.S.-meeting and 7-th Prague Conference on Information Theory, Statistical decision functions and Random processes.