Adaptive control of specially structured Markov chains

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 76-28

Adaptive control of specially structured Markov chains

by

K.M. van Hee

Eindhoven, December 1976

The Netherlands

Adaptive control of specially structured Markov chains

by

K.M. van Hee

0. Summary

We consider Markov decision processes where the state at time $n+1$ is a function of the state at time n, the action at time n and the outcome of a random variable Y_{n+1}. The random variables $Y_1, Y_2, Y_3, \ldots$ are independent and identically distributed with an incompletely known distribution. The class of problems considered includes the linear system with quadratic cost and a simple inventory control model. The minimal Bayesian expected total cost is determined or approximated. The strategy that takes, at each time, the action that is optimal if the estimated distribution is the true distribution, is studied.

1. Introduction and preliminaries

Consider a Markov decision process with *state space* $X \subset \mathbb{R}^{N_1}$ and *action space* $D \subset \mathbb{R}^{N_2}$. The *cost function* $k: X \times D \to \mathbb{R}$ is Borel measurable and bounded from below. The state of the system at time n, X_n is determined by a measurable function F:

$$X_n = F(X_{n-1}, U_{n-1}, Y_n), \quad n = 1,2,3,\ldots$$

where U_{n-1} is the action at time $n-1$ and $\{Y_n, n = 1,2,3,\ldots\}$ are independent and identically distributed random variables in $\mathbb{R}^{N_3}$, not *controllable* by the decisionmaker. At time n Y_n becomes *visible* to him. The process $\{Y_n, n = 1,2,3,\ldots\}$ is called the *external process*. The distribution of Y_n is not completely known: Y_n has a probability density $p(\cdot|\theta)$ with respect to a σ-finite measure m where θ is the unknown *parameter* belonging to the *parameter space* Θ, a completely separable metric space endowed with the Borel σ-field H. Let Π denote the set of all *strategies* which are based on the *visible histories*, i.e. for $\pi \in \Pi$ the action U_n may depend on $X_0,\ldots,X_n,U_0,\ldots$ $\ldots,U_{n-1},Y_1,\ldots,Y_n$ (see van Hee (1976A) for a formal definition).
For each $x \in X$, $\pi \in \Pi$ and $\theta \in \Theta$ we have a random process $\{(X_n,U_n,Y_{n+1}), n = 0,1,2,\ldots\}$ and a probability measure $\mathbb{P}^{\pi}_{x,\theta}$ on the sample space of the process. (The expectation with respect to this probability is denoted by $\mathbb{E}^{\pi}_{x,\theta}$.)

Future cost are discounted by $\beta \in [0,1)$. The *expected total cost* $v(x,\theta,\pi)$, $x \in X$, $\theta \in \Theta$, $\pi \in \Pi$ is defined by

$$v(x,\theta,\pi) := \mathbb{E}^{\pi}_{x,\theta}[\sum_{n=0}^{\infty} \beta^n k(X_n,U_n)] .$$

We assume that for each $y \in \mathbb{R}^{N_3}$ $p(y|\cdot)$ is H-measurable. Let W be the set of all probability measures on (Θ,H) and let W be the Borel σ-field generated by the weak topology on W. We identify each $\theta \in \Theta$ with the distribution in W that is degenerated in θ.
In the Bayesian approach we fix $q \in W$ and we assume that the parameter θ is a random variable Z with *prior distribution* q on Θ. After observing $Y_1,\ldots,Y_n$ we have the *posterior distribution* Q_n on W:

1.1. $$Q_n(B) := \mathbb{P}_q[Z \in B \mid Y_1,\ldots,Y_n],\ B \in H,\ n = 1,2,3,\ldots$$
$$Q_0 := q$$

where $\mathbb{P}_q$ is defined by

$$\mathbb{P}_q[Z \in B,\ Y_1 \in C_1,\ldots,Y_n \in C_n] := \int_B q(d\theta)\mathbb{P}_\theta[Y_1 \in C_1,\ldots,Y_n \in C_n]$$

for $B \in H$ and C_i a Borel subset of $\mathbb{R}^{N_3}$, $i = 1,2,\ldots,n$. (Note that we write $\mathbb{P}_\theta$ instead of $\mathbb{P}^{\pi}_{x,\theta}$ when we are dealing with the external process or the Bayes process. Sometimes we use $\mathbb{P}[\ \mid Q_0 = q] := \mathbb{P}_q[\ ]$).

We call the process $\{Q_n,\ n = 0,1,2,\ldots\}$ the *Bayes process*. We first introduce some notations: for $y \in \mathbb{R}^{N_3}$, $\varphi \in W$

1.2. $$p(y,\varphi) := \int p(y|\theta)\varphi(d\theta)$$

and if $p(y,\varphi) > 0$:

1.3. $$T_y(\varphi)(B) := \int_B p(y|\theta)\varphi(d\theta).\{p(y,\varphi)\}^{-1},\ B \in H .$$

Assume that for all $\varphi \in W$ there is an *stationary optimal strategy* if $p(\cdot,\varphi)$ is the density of the external process; i.e. there is for each $\varphi \in W$ a function f_φ: $X \to D$ such that it is optimal to choose $U_n = f_\varphi(X_n)$, $n = 0,1,2,\ldots$

To control the process when the parameter is unknown one could use the strategy, given by $U_n = f_{Q_n}(X_n)$. We call this strategy the *Bayesian equivalent rule*. In fact $p(\cdot,Q_n)$ is the Bayes estimator of the density of the external process at time n. If the controller uses the Bayesian equivalent rule, he determines at time n: $p(\cdot,Q_n)$ and the optimal control for the model with this density. Then he uses this control for one time period. In [Mandl (1974)] this strategy is examined with respect to the average cost criterion and more general estimation procedures. In this paper we study this strategy with respect to the *Bayesian expected total cost:*

$$v(x,q,\pi) := \int v(x,\theta,\pi)q(d\theta) .$$

We show that for the linear system with quadratic cost the Bayesian equivalent rule is optimal (section 2) and also for models where k is separable, i.e. $k(x,u) = a(x) + b(u)$ and where $F(x,u,y)$ does not depend on x (section 3). Finally we consider in section 3 a simple inventory control model (without fixed order cost) and we give approximations for the value of the Bayesian equivalent rule.

We conclude this section with some preparations. We consider the Bayesian decision problem for all prior distributions $q \in W$ simultaneously. The value function $v: X \times W \to \mathbb{R}$ is defined by

1.4. $$v(x,q) := \inf_{\pi\in\Pi} v(x,q,\pi) .$$

The Bayesian decision problem can be reduced to a *dynamic program* with state space $X \times W$, action space D and costfunction $k(x,q,u) := k(x,u)$. See [Rieder (1975)] for a proof of this statement if $F(x,u,\cdot)$ is a one to one mapping and in [Van Hee (1976A)] this is proved for the general situation in a similar way.

For this dynamic program we define the standard operators: Let $g: X \times W \to \mathbb{R}$ be such that the following expression is defined for all $f \in \tilde{D}$

1.5. $$(L_f g)(x,q) := k(x,f(x,q)) + \beta \int g(F(x,f(x,q),y),T_y(q))p(y,q)m(dy) \quad ,$$

where $\tilde{D} := \{f: X \times W \to D \mid f \text{ measurable}\}$

1.6. $$(Ug)(x,q) := \inf_{f\in\tilde{D}} (L_f g)(x,q) .$$

A strategy $\pi \in \Pi$ such that $U_n = f(X_n, Q_n)$ for all $n = 0,1,2,\dots$ is called *stationary*, if $f \in \tilde{D}$.
For each $q \in W$ the Bayes process forms a (stationary) *Markov chain* and if the right-hand side is defined we have:

$$Q_{n+1} = T_{Y_{n+1}}(Q_n)$$

(see [Van Hee (1976A)]).

Lemma 1.1. Let $f\colon \Theta \to \mathbb{R}$ be bounded and measurable. We extend f to a function on W by

$$f(q) := \int f(\theta)q(d\theta),\ q \in W .$$

For $m \le n$ it holds that $\mathbb{E}[f(Q_n) \mid Q_m] = f(Q_m)$.

Proof. First let

$$f(\theta) := \sum_{k=1}^{N} a_k 1_{A_k}(\theta),\ A_k \in H,\ k = 1,\dots,N .$$

Then it holds that

$$f(q) = \sum_{k=1}^{N} a_k q(A_k)$$

and

$$\mathbb{E}[f(Q_n) \mid Q_m] = \sum_{k=1}^{N} a_k . \mathbb{E}[Q_n(A_k) \mid Q_m] = \sum_{k=1}^{N} a_k Q_m(A_k)$$

(see [Van Hee (1976A)] for the last equality).
Hence the statement is verified for step functions. Using standard arguments it is easy to derive the desired result. □

2. Linear systems with quadratic cost

In this section we use ideas and concepts which are familiar in the theory of linear systems (see [Kushner (1971), chpt. 9]). The model specifications are as follows.
The state space $X = \mathbb{R}^N$, the action space $D = \mathbb{R}^M$ the external process takes on values in $\mathbb{R}^N$. The cost function is defined by

$$k(x,u) := x'Rx + u'Su$$

where R is a nonnegative definite $N \times N$ matrix and S a positive definite

$M \times M$ matrix (x' is the transpose of x). The transition mechanism is given by $F(x,u,y) := Ax + Bu + y$ where the $N \times N$ matrix A and the $N \times M$ matrix B satisfy the controllability assumption:

2.1. $\quad \operatorname{rank}[B, AB, \ldots, A^{N-1}B] = N$.

For $q \in W$ we define the vector μ_q and the matrices M_q and Σ_q:

$$\mu_q(i) := \int y_i p(y,q)m(dy); \; M_q(i,j) := \int \mu_\theta(i)\mu_\theta(j)q(d\theta);$$

$$\Sigma_q(i,j) := \int y_i y_j p(y,q)m(dy)$$

(for $y \in \mathbb{R}^N$ y_i is the i-th component of y). Note that $\Sigma_q - M_q$ is the covariance matrix of Y_n averaged over θ with q. Throughout this section we assume that

2.2. $\quad \int |y_i y_j| p(y|\theta)m(dy)$

is bounded over θ. Hence, μ_q, M_q and Σ_q are bounded on W.

Lemma 2.1. For $q \in W$ it holds that

i) $\quad \int \mu_{T_y(q)}(i)p(y,q)m(dy) = \mu_q(i)$.

ii) $\quad \int y_j \mu_{T_y(q)}(i)p(y,q)m(dy) = M_q(i,j)$.

Proof.

$$\mu_{T_y(q)}(i)p(y,q) = p(y,q) \int z_i \{\int \frac{p(z|\theta)p(y|\theta)}{p(y,q)} q(d\theta)\}m(dz) =$$

$$= \int \{\int z_i p(z|\theta)p(y|\theta)m(dz)\}q(d\theta) .$$

Hence

$$\int \mu_{T_y(q)}(i)p(y,q)m(dy) = \int \{\int z_i p(z|\theta)m(dz)\}q(d\theta) = \mu_q(i)$$

and

$$\int y_j \mu_{T_y(q)}(i)p(y,q)m(dy) = \iiint y_j z_i p(z|\theta)p(y|\theta)m(dz)m(dy)q(d\theta) =$$

$$= \int \{\int y_j p(y|\theta)m(dy)\}.\{\int z_i p(z|\theta)m(dz)\}q(d\theta) = M_q(i,j) .$$

(Note that all changings of integration order are allowed by 2.2.) □

The next lemma describes the behavior of the U-operator defined in 1.6. The proof proceeds in a familiar way (see [Kushner (1971), section 9.2.2]).

Lemma 2.2. Let

$$f(x,q) := x'Px + x'L\mu_q + H(q),\ x \in X,\ q \in W\ ,$$

where P is a nonnegative definite matrix, L a N × N matrix and H a bounded continuous function on W, then

$$(Uf)(x,q) = x'\tilde{P}x + x'\tilde{L}\mu_q + \tilde{H}(q),\ x \in X,\ q \in W$$

where

$$\tilde{P} := F_1(P) := R + \beta A'PA - \beta^2 A'PB(S + \beta B'PB)^{-1}B'PA$$

2.3. $$\tilde{L} := F_2(L,P) := 2\beta A'P + \beta A'L - \beta^2 A'PB(S + \beta B'PB)^{-1}(2B'P + B'L)$$

$$\tilde{H}(q) := F_3(H,q,P,L) := -\tfrac{1}{4}\beta^2\mu_q'(2PB + L'B)(S + \beta B'PB)^{-1}(2B'P + B'L)\mu_q$$
$$+ \beta \int H(T_y(q))p(y,q)m(dy) + \beta\ \mathrm{trace}(P\Sigma_q) + \beta\ \mathrm{trace}(LM_q)\ .$$

And the minimizing control u(x,q) is

$$u(x,q) = -\beta(S + \beta B'PB)^{-1}B'PAx - \beta(S + \beta B'PB)^{-1}\{B'P + \tfrac{1}{2}B'L\}\mu_q\ .$$

Remark. Note that $F_3(H,\cdot,P,L)$ is bounded and continuous function on W since μ_q, Σ_q and M_q are and because $T_y(\cdot)$ is continuous.

Proof. By some evaluations, using lemma 2.1 we get

$$(Uf)(x,q) = \inf_u\{u'(S + \beta B'PB)u + (2\beta x'A'PB + 2\beta\mu_q'PB + \beta\mu_q'L'B)u\} +$$
$$+ x'(R + \beta A'PA)x + \beta x'(2A'P + A'L)\mu_q +$$
$$+ \beta \int H(T_y(q))p(y,q)m(dy) + \beta\ \mathrm{trace}(P\mu_q) + \beta\ \mathrm{trace}(LM_q)\ .$$

Since P is nonnegative definite and S is positive definite we have the existence of $(S + \beta B'PB)^{-1}$. Hence by a standard argument for the minimization of quadratic forms we have the desired result. □

Now we shall consider the sequence of *successive approximations* $v_n(x,q) := (U^n 0)(x,q)$ and we define sequences of N × N matrices $\{P_n,\ n = 0,1,2,\ldots\}$, $\{L_n,\ n = 0,1,2,\ldots\}$ and a sequence of bounded continuous functions on W: $\{H_n,\ n = 0,1,2,\ldots\}$, $P_0 := 0$, $L_0 := 0$, $H_0 := 0$ and for $n = 0,1,2,\ldots$

$$P_{n+1} := \mathcal{F}_1(P_n)$$

2.4. $$L_{n+1} := \mathcal{F}_2(L_n, P_n)$$

$$H_{n+1} := \mathcal{F}_3(H_n, \cdot, P_n, L_n) \ .$$

It is a direct consequence of lemma 2.2, that

$$v_n(x,q) = x'P_n x + x'L_n\mu_q + H_n(q) \ .$$

In lemma 2.3 we prove that P_n converges, elementwise, to a nonnegative definite matrix P^* and that L_n converges to matrix L^*. The proof of $P_n \to P^*$ can also be found in [Kushner, 1971, section 9.2.3].

Lemma 2.3.

i) P_n converges to a nonnegative definite matrix P^*, satisfying $P^* = \mathcal{F}_1(P^*)$.
ii) L_n converges to a matrix L^*, satisfying $L^* = \mathcal{F}_2(L^*, P^*)$.

Proof. Since P_n and L_n do not depend on the external process their limiting behavior is the same if we assume $Y := \mu \in \mathbb{R}^N$, i.e. Y has a degenerated distribution in μ for all $\theta \in \Theta$. Now we have a deterministic linear system. The value of this system is denoted by $v(x)$ and the sequence of successive approximations by $v_n(x)$. First we show that for this system the value is finite. Let $x = x_0$ be the starting state. Note that

$$x_N = A^{N-1}x_0 + \sum_{k=0}^{N-1} A^k B u_{N-1-k} + \sum_{k=0}^{N-1} A^k \mu$$

hence

$$x_N - A^{N-1}x_0 - \sum_{k=0}^{N-1} A^k \mu = \sum_{k=0}^{N-1} A^k B u_{N-1-k} \ .$$

By the controllability assumption 2.1 we may choose actions $u_0,\ldots,u_{N-1}$ such that $x_N = 0$ and so there is a strategy π such that $x_{kN} = 0$ for $k = 1,2,3,\ldots$ Since we have a discount factor $0 < \beta < 1$ we see that the total cost of π is finite. Hence $v_n(x)$ is bounded in n, and so we have $v_n(x)$ converges for each x. Note that

$$v_n(x) = x'P_n x + x'L_n\mu + H_n$$

where H_n is defined in 2.3 and 2.4 for this special external process. Note that H_n does not depend on x.

If $\mu = 0$ we have $x'P_nx$ converges for all x, which implies that P_n converges (elementwise). Since $v_n(0)$ converges we have that H_n converges. So $x'L_n\mu$ converges for all x and μ. Therefore L_n converges. Since F_1 and F_2 are continuous functions elementwise, we have $F_1(P^*) = P^*$ and $F_2(L^*,P^*) = L^*$. □

In lemma 2.4 we show that $H_n(q)$ converges in general.

Lemma 2.4. H_n converges to a bounded and continuous function H^* satisfying

$$F_3(H^*,\cdot,P^*,L^*) = H^*(\cdot) .$$

Proof. Let

$$b_n(q) := -\tfrac{1}{4}\beta^2\mu_q'(2P_nB + L_n'B)(S + \beta B'P_nB)^{-1}(2B'P_n + B'L_n)\mu_q$$

$$+ \beta\ \mathrm{trace}(P_n\Sigma_q) + \beta\ \mathrm{trace}(L_nM_q) .$$

Note that $b_n(q)$ converges and call $b(q) := \lim_{n\to\infty} b_n(q)$. We have, in terms of the Bayes process:

$$H_{n+1}(q) = b_n(q) + \beta \int H_n(T_y(q))p(y,q)m(dy) =$$

$$= b_n(q) + \beta\mathbf{E}[H_n(Q_1) \mid Q_0 = q] .$$

Iterating this equation yields

$$H_{n+1}(q) = \sum_{k=0}^{n} \beta^k\mathbf{E}[b_{n-k}(Q_k) \mid Q_0 = q]$$

since the Bayes process is a Markov chain and $H_0 = 0$. Note that $b_n(q)$, as function of n and q, is bounded since P_n, L_n and μ_q are. Hence for all $\varepsilon > 0$ there is a N such that

$$\sum_{k=N+1}^{n} \beta^k\mathbf{E}[b_{n-k}(Q_k) \mid Q_0 = q] < \varepsilon \quad (n \geq N+1) .$$

By the dominated convergence theorem we have for all k

$$\lim_{n\to\infty} \mathbf{E}[b_{n-k}(Q_k) \mid Q_0 = q] = \mathbf{E}[b(Q_k) \mid Q_0 = q] .$$

Hence

$$\lim_{n\to\infty} H_n(q) = \sum_{k=0}^{\infty} \beta^k\mathbf{E}[b(Q_k) \mid Q_0 = q] =: H^*(q) .$$

It is easy to verify that $H^*(q) = b(q) + \beta \mathbb{E}[H^*(Q_1) \mid Q_0 = q]$ which shows that $H^* = F_3(H^*,\cdot,P^*,L^*)$. □

We resume the following definitions, given in lemma 2.4:

2.5.
$$b(q) := -\tfrac{1}{4}\beta^2 \mu_q'(2P^*B + L^*B)(S + \beta B'P^*B)^{-1}(2B'P^* + B'L^*)\mu_q + \beta\,\mathrm{trace}(P^*\Sigma_q) + \beta\,\mathrm{trace}(L^*M_q)$$

$$H^*(q) := \sum_{n=0}^{\infty} \beta^n \mathbb{E}[b(Q_n) \mid Q_0 = q] .$$

The next theorem is one of the main results of this section. It gives an explicit expression for the optimal strategy and for the value function.
In fact the optimal strategy is a *linear control* (see [Kushner (1971)]) and it is a Bayesian equivalent rule also.

Theorem 2.5.

i) The value function satisfies

$$v(x,q) = x'P^*x + x'L^*\mu_q + H^*(q) .$$

ii) The optimal strategy chooses in state (x,q) the action

$$u(x,q) = -\beta(S + \beta B'P^*B)^{-1}B'P^*Ax - \beta(S + \beta B'P^*B)^{-1}(B'P^* + \tfrac{1}{2}B'L^*)\mu_q ,$$

(where P^* and L^* are defined in lemma 2.3).

Proof. It follows from lemmas 2.2, 2.3 and 2.4 that

$$v_\infty(x,q) := \lim_{n\to\infty} v_n(x,q) = x'P^*x + x'L^*\mu_q + H^*(q) ,$$

and also

$$v_\infty(x,q) = (Uv_\infty)(x,q) = (L_u v_\infty)(x,q)$$

where u represents the stationary strategy defined in 2.6. Hence by [Schäl (1975), thm. 5.3.1] we have the desired result. □

In the next theorem we compare the value of our Bayesian control model with the values of two other models.
First we consider the model where the parameter θ is chosen according to the probability q, but before the controller starts to control the system he will be informed about the chosen value θ. Hence his expected total cost will be:

$$\int v(x,\theta)q(d\theta).$$

On the other hand we consider the model with a completely known external process with probability density

$$\int p(\cdot|\theta)q(d\theta)$$

with respect to m. We call the value of this process w(x,q). With these processes we can give bounds for the extra cost we have by the lack of information over the parameter.

Theorem 2.6.

i) $$\int v(x,\theta)q(d\theta) \le v(x,q) \le w(x,q)$$

ii) $$\frac{1}{1-\beta}\int b(\theta)q(d\theta) \le H(q) \le \frac{b(q)}{1-\beta}$$

iii) $$v(x,q) - \int v(x,\theta)q(d\theta) \le \frac{1}{1-\beta}\{b(q) - \int b(\theta)q(d\theta)\} .$$

Proof. Since

$$v(x,q) = \inf_{\pi\in\Pi} \int v(x,\theta,\pi)q(d\theta) \ge \int \inf_{\pi\in\Pi} v(x,\theta,\pi)q(d\theta) = \int v(x,\theta)q(d\theta).$$

The left-hand side of has been proved. Note that

$$G := (2P^*B + L^*B)(S + \beta B'P^*B)^{-1}(2B'P^* + B'L^*)$$

is positive definite since $(S + \beta B'P^*B)$ is. Hence G can be written as $C'\Lambda C$ where C is orthogonal and Λ is a diagonal matrix with nonnegative entries $\lambda_1,\dots,\lambda_N$. And therefore

$$\mu_q' G\mu_q = \sum_{i=1}^N \lambda_i\{\sum_{j=1}^N c_{ij}\mu_q(i)\}^2 .$$

Hence, by Jensen's inequality:

$$\mathbf{E}_q[\mu_{Q_n}' G\mu_{Q_n}] \ge \sum_{i=1}^N \lambda_i[\mathbf{E}_q\{\sum_{j=1}^N c_{ij}\mu_{Q_n}(j)]\}^2$$

and by lemma 1.1

$$\mathbf{E}_q[\mu_{Q_n}' G\mu_{Q_n}] \ge \sum_{i=1}^N \lambda_i\{\sum_{j=1}^N c_{ij}\mu_q\}^2 .$$

Note that

$$\mathrm{trace}(L^* M_q) = \sum_{i=1}^{N} \sum_{j=1}^{N} L^*(i,j) \int \mu_\theta(i)\mu_\theta(j)q(d\theta)$$

and that

$$\mathrm{trace}(P^* \Sigma_q) = \sum_{i=1}^{N} \sum_{j=1}^{N} P^*(i,j) \int \{\int y_i y_j p(y|\theta)m(dy)\} q(d\theta) .$$

Hence by lemma 1.1

$$\mathbf{E}_q[\mathrm{trace}(L^* M_{Q_n})] = \mathrm{trace}(L^* M_q) \text{ and } \mathbf{E}_q[\mathrm{trace}(P^* \Sigma_{Q_n})] = \mathrm{trace}(P^* \Sigma_q) .$$

Therefore we have $\mathbf{E}_q[b(Q_n)] \le b(q)$. It is easy to verify that

$$w(x,q) = x'P^*x + x'L^*\mu_q + \frac{b(q)}{1-\beta}$$

and that

$$\int v(x,\theta)q(d\theta) = x'P^*x + x'L^*\mu_q + \frac{\int b(\theta)q(d\theta)}{1-\beta} .$$

This implies the assertions of the theorem. □

3. Bayesian equivalent rules and a simple inventory model

In this section we consider an adaptive control problem with the property that the Bayesian equivalent rule is optimal. We apply results for this model to a simple inventory control problem afterwards.

The model we are dealing with here is specified by:

3.1. i) D is compact.

ii) $k(x,u) := a(x) + b(u)$ where a and b are lower semi continuous and a is bounded from below.

iii) the transition function $F(x,u,y)$ does not depend on the first coordinate and is continuous in the second. (We shall write $F(u,y)$ instead of $F(x,u,y)$.)

iv) $\int a(F(u,y))p(y|\theta)m(dy)$ is bounded over Θ for all $u \in D$.

It is easy to verify that this model satisfies the conditions C and W of [Schäl (1975)] which implies that:

3.2. i) $v_n(x,q) := (U^n 0)(x,q)$ converges to the value function v (pointwise).

ii) There is an optimal stationary strategy.

Theorem 3.1. The value function v of the model given by 3.1 satisfies

$$v(x,q) = a(x) + \sum_{n=0}^{\infty} \beta^n \mathbb{E}[d(Q_n) \mid Q_0 = q]$$

where

$$d(q) := \inf_{u \in D}\{b(u) + \beta \int a(F(u,y))p(y,q)m(dy)\}$$

is bounded and continuous on W. The following holds: there is a measurable function s: $W \to D$ such that the optimal strategy chooses in state (x,q) the action $u(x,q) = s(q)$. Since

Proof. Since

$$b(u) + \int a(F(u,y))p(y,q)m(dy)$$

is lower semi continuous and 3.1iv) we have that d is bounded and continuous on W. Let $e := \min_{u \in D}\{b(u)\}$. Then since $v_0(x,q) = 0$ for all $x \in X$, $q \in W$ we have $v_1(x,q) = a(x) + e$. With induction we prove that

$$(*) \qquad v_n(x,q) = a(x) + \sum_{k=0}^{n-2} \beta^k \mathbb{E}[d(Q_k) \mid Q_0 = q] + \beta^{n-1} e .$$

Assume (*) holds for n. Then

$$v_{n+1}(x,q) = a(x) + d(q) + \beta \int \sum_{k=0}^{n-2} \beta^k \mathbb{E}[d(Q_k) \mid Q_0 = T_y(q)]p(y,q)m(dy) + \beta^n e.$$

Using the Markov property of the Bayes process we have the assertion. By 3.2 we have an optimal stationary strategy and by considering the optimality equation it is easy to see that the optimal action in (x,q) can be chosen independently of x. □

Remarks.

1. The optimal strategy is a myopic rule since the optimal strategy for the n-horizon problem is the same for all $n \geq 2$.
2. Note that v(x,q) is a separable function, i.e. $v(x,q) = a(x) + h(q)$ where

$$h(q) := \mathbb{E}[\sum_{n=0}^{\infty} \beta^n d(Q_n) \mid Q_0 = q] .$$

   In fact this property guarantees that the Bayes equivalent rule is optimal in more general models.
3. We call s(q) the *control point*.

In the next theorem we have bounds for the value function in a way similarly to theorem 2.6.

Theorem 3.2.

$$a(x) + (1-\beta)^{-1} \int d(\theta)q(d\theta) \le v(x,q) \le a(x) + (1-\beta)^{-1} d(q) .$$

Proof. Since

$$\mathbf{E}_q[d(Q_n)] \le \inf_{u \in D} \mathbf{E}_q[b(u) + \beta \int a(F(u,y))p(y,Q_n)m(dy)]$$

we have by lemma 1.1

$$\mathbf{E}_q[d(Q_n)] \le \inf_{u \in D}\{b(u) + \beta \int a(F(u,y))p(y,q)m(dy)\} = d(q) .$$

This gives the right-hand inequality; the left-hand side proceeds analogously to theorem 2.6i). □

Now we shall consider an inventory control model which is narrowly related to models of the type described in 3.1: the only difference is that the actions allowed in state x depend on x.
We call this model (A). Interesting results for this model are given by [Scarf (1959)], [Iglehart (1964)] and [Rieder (1972)].

Model (A):

i) $X := \{x \in \mathbf{R} \mid x \le M\}$, $M > 0$ is the *capacity*.
ii) $D_x := \{u \in \mathbf{R} \mid x \le u \le M\}$, u is the *inventory* after ordering.
iii) the external process is one dimensional and represents the *demand:* $p(y|\theta) = 0$ for $y \le 0$ for all $\theta \in \Theta$ and $\sup_{\theta\in\Theta} \mu_\theta < \infty$.
iv) $k(x,u) := hx^+ + px^- + c(u-x)$ where h is the *holding cost*, p the *shortage cost* and c the *production cost*, $h,p,c > 0$ and $\beta(p+c) > c$.
v) $F(x,u,y) := u - y$, $u \in D_x$.

We call the value function of model (A): v. We shall compare model (A) with model (B), which model only differs from (A) by its action space:

Model (B): $D := \{u \in \mathbf{R} \mid 0 \le u \le M\}$, further specifications as in model (A). The value function for model (B) will be denoted by w. The control point s(q) for model (B) can be chosen as the minimum of M and the smallest $u \ge 0$ such that

3.3. $$\lim_{\varepsilon\downarrow 0} \int_0^{u-\varepsilon} p(y,q)m(dy) \le \frac{p - \frac{1-\beta}{\beta} c}{p+n} \le \int_0^u p(y,q)m(dy) .$$

Note that $s(q) > 0$ for all $q \in W$, since $\beta(p+c) > c$. We shall consider for model (A) the strategy that orders until $s(q)$ if possible, i.e.

3.4. $$u(x,q) := \max\{x, s(q)\}$$

the value of this strategy is denoted by $\hat{v}$.

If we are dealing with a known parameter θ this strategy $u(x,\theta) = s(\theta)$ is optimal for model (A), and it is the Bayesian equivalent rule for the adaptive control of model (A). It is our goal to compare v, w and $\hat{v}$. First we need some preparations.

Lemma 3.3.

i) There is a measurable function $t\colon W \to X$ such that there is an optimal strategy for model (A) satisfying:

$$u(x,q) = \max\{x, t(q)\} .$$

ii) The control point $s(q)$ for model (B) satisfies

$$s(q) \ge t(q) \quad \text{for all } q \in W .$$

Proof.

i) See [Rieder (1972), th. 7.2 and th. 7.3].

ii) Let $f(x,q) := v(x,q) - \{hx^+ + px^- - cx\}$.

By the optimality equation for model (A) we have

$$f(x,q) = \inf_{M \ge u \ge x} \{cu + \beta \int v(u-y, T_y(q)) p(y,q) m(dy)\}.$$

Therefore $f(\cdot,q)$ is nondecreasing for all $q \in W$. Note that f satisfies:

(*) $$f(x,q) = \inf_{x \le u \le M} \{cu + \beta \int [h(u-y)^+ + p(u-y)^- - c(u-y)] p(y,q) m(dy) + \beta \int f(u-y, T_y(q)) p(y,q) m(dy)\} .$$

Note that, by considering model (B),

$$cu + \beta \int [h(u-y)^+ + p(u-y)^- - c(u-y)] p(y,q) m(dy)$$

is convex and attains a minimum on $[0,M]$ in $s(q)$ and note further that

$$\beta \int f(u-y), T_y(q))p(y,q)m(dy)$$

is nondecreasing. Hence the minimizer of $(*)$, $t(q)$, must satisfy $t(q) \le s(q)$. □

Lemma 3.4. For each strategy π for model (A), which has the property that $0 \le U_n \le M$ it holds that for some $\Delta > 0$:

$$v(x,q,\pi) \le hx^+ + px^- - cx + \Delta \ .$$

Proof.

$$v(x,q,\pi) \le hx^+ + px^- - c(M-x) + \sum_{n=1}^{\infty} \beta^n \int q(d\theta)\mathbb{E}_\theta[h(M-Y_n)^+ + p(0-Y_n)^- + c(M-Y_n))$$

$$\le hx^+ + px^- - cx + cM + \frac{1}{1-\beta}\{(h+p-c)\mu_q + (c+h)M\} \ .$$ □

Lemma 3.5. It holds that

$$\hat{v}(x+\Delta,q) - \hat{v}(x,q) \le \frac{h\Delta}{1-\beta} \quad \text{for all } \Delta > 0,\ q \in W \ .$$

Proof. Let X_n denote the inventory at time n using the control 3.4 if the starting state is x and $\tilde{X}_n$ if the starting state is $x+\Delta$. Note that X_n and $\tilde{X}_n$ both satisfy the recurrence relation in z:

$$z_{n+1} = \max\{z_n, s(Q_n)\} - Y_{n+1} \ .$$

Hence

$$0 \le \tilde{X}_n - X_n \le \Delta \text{ for } n = 0,1,2,\dots \ .$$

And the difference between the direct cost for both processes at time n:

$$h(\tilde{X}_n^+ - X_n^+) + p(\tilde{X}_n^- - X_n^-) + c\{(s(Q_n) - \tilde{X}_n)^+ - (s(Q_n) - X_n)^-\} \le h\Delta.$$

This proves the lemma. □

In the following theorem we give bounds for the difference of the value functions for models (A) and (B). Define:

$$S_n := s(Q_n),\ n = 0,1,2,\dots \ .$$

Theorem 3.6. For all $x \in X$, $q \in W$ we have

i) $\quad w(x,q) \le v(x,q) \le \hat{v}(x,q) \ .$

ii) $$\hat{v}(x,q) - w(x,q) \le \{\frac{\beta}{1-\beta} h + c\}\{(x - s(q))^+ + \sum_{n=1}^{\infty} \beta^n \mathbf{E}_q[(S_{n-1} - Y_n - S_n)^+]\}.$$

Proof.

i) Note that the lower bound for the action space in model (B) is not essential, hence $w(x,q) \le v(x,q)$.

ii) Define $\ell(x,q) := \hat{v}(x,q) - w(x,q)$.

For $x \le s(q)$ we have

(*) $$\ell(x,q) = \beta \int \ell(s(q) - y, T_y(q))p(y,q)m(dy) = \ell(s(q),q) .$$

For $x > s(q)$ we have by lemma 3.5

$$\hat{v}(x,q) \le \hat{v}(s(q),q) + (x - s(q))\frac{h}{1-\beta} .$$

And therefore, since

$$w(x,q) = w(s(q),q) + (h - c)(x - s(q))$$

it holds that

(**) $$\ell(x,q) \le \ell(s(q),q) + (x - s(q)) + \{\frac{\beta}{1-\beta} h + c\} .$$

Let $A := \frac{\beta}{1-\beta} h + c$. By (*) and (**) we have in terms of the Bayes process:

$$\ell(x,q) \le A(x - s(q))^+ + \beta\mathbf{E}[\ell(S_0 - Y_1, Q_1) \mid Q_0 = q] .$$

And since the Bayes process forms a Markov chain, for $n = 0,1,2,...$

$$\ell(S_n - Y_{n+1}, Q_{n+1}) \le A(S_n - Y_{n+1} - S_{n+1})^+ + \beta\mathbf{E}[\ell(S_{n+1} - Y_{n+2}, Q_{n+2}) | Q_{n+1}].$$

Iterating this equation yields:

(***) $$\mathbf{E}[\ell(S_0 - Y_1, Q_1) | Q_0 = q] \le A \sum_{k=1}^{N} \beta^k \mathbf{E}[(S_n - Y_{n+1} - S_{n+1})^+ | Q_0 = q] + \beta^{N+1}\mathbf{E}[\ell(S_{N+1} - Y_{N+2}, Q_{N+2}) | Q_0 = q] .$$

Let $d(q) := w(x,q) - \{hx^+ + px^- - cx\}$ (note that d does not depend on x). Then by lemma 3.4 and theorem 3.2 we have for some $\Delta > 0$:

$$0 \le \ell(x,q) \le \Delta - \frac{\int d(\theta)q(\theta)}{1-\beta} \le \Delta - \inf_{\theta\in\Theta} d(\theta)\{1-\beta\}^{-1} < \infty .$$

Hence the last term of (***) tends to zero if N tends to infinity. □

Corollary 3.7. If for all $q \in W$ it holds that

$$3.5. \qquad \int_{\{y \mid s(q)-y \le s(T_y(q))\}} p(y,q)m(dy) = 1$$

then for all $x \le s(q)$ we have $v(x,q) = w(x,q)$ and therefore the Bayesian equivalent rule is optimal.

We conclude this section with some remarks:

1) The statement of corollary 3.7 is not new. In [Veinott (1965) section 6] a similar condition is considered for a multiproduct inventory model with dependent demand to prove an analogous statement. In [Rieder (1972), th. 7.6] Veinott's result is proved in the Bayesian inventory problem. The inequality of theorem 3.6 ii) seems to be new and it gives us the opportunity to compute an upper bound for the value belonging to the Bayesian equivalent rule.
2) The condition 3.5 is fulfilled in the following situation. Let

$$G(u,q) := \int_0^u p(y,q)m(dy)$$

and assume that $G(\cdot,\theta)$ is continuous for all $\theta \in \Theta$. The control point $s(q)$ is the smallest root of

$$G(u,q) = \frac{p - \frac{1-\beta}{\beta} c}{p + h} .$$

Define

$$s_{min} := \inf_{\theta \in \Theta} s(\theta), \quad s_{max} := \sup_{\theta \in \Theta} s(\theta) .$$

It is easy to verify that $s_{min} \le s(q) \le s_{max}$ for all $q \in W$. If there is an $e > 0$ such that

i) $G(e,\theta) > 0$ for all $\theta \in \Theta$.
ii) $e \ge s_{max} - s_{min}$

then 3.5 holds.
3) In [van Hee (1976B)] methods are studied to approximate the value of a Bayesian control problem in case where X, D and Θ are finite. If we are dealing with models, which approximate the structure of models given in 3.1 then the approximation methods are very good.

Literature

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

2. Van Hee, K.M. (1976B) Approximations in Bayesian Controlled Markov Chains, to appear in: the Proceedings of the advanced seminar on Markov decision theory in Amsterdam, in the series of Mathematical Centre Tracts.

3. Iglehart, D.L. (1964) The dynamic inventory problem with unknown demand distribution, Management Science 10, 429-440.

4. Kushner, H. (1971) Introduction to stochastic control. Holt, Rinehart and Winston, Inc.

5. Mandl, P. (1974) Estimation and Control in Markov Chains. Adv. Appl. Prob. 6, 40-60.

6. Rieder, U. (1972) Bayessche dynamische Entscheidungs- und Stoppmodelle, dissertation, Hamburg.

7. Rieder, U. (1975) Bayesian Dynamic Programming. Adv. Appl. Prob. 7, 330-348.

8. Scarf, H. (1959) Bayes solution of the statistical inventory problem. Ann. Math. Stat. 30, 490-508.

9. Schäl, M. (1975) Conditions for Optimality in Dynamic Programming and for the Limit of n-stage optimal policies to be optimal. Z. Wahrscheinlichkeitstheorie verw. Gebiete 32, 179-196.