Control and Cybernetics

vol. **42** (2013) No. 3

Optimal stopping model with unknown transition probabilities[†]

by

Masayuki Horiguchi[1] and Alexey B. Piunovskiy[2]

[1]Department of Mathematics, Faculty of Science
Kanagawa University
2946 Tsuchiya, Hiratsuka-shi, Kanagawa 259-1293, Japan
horiguchi@kanagawa-u.ac.jp
[2]Department of Mathematical Sciences
University of Liverpool
L69 7ZL, Liverpool, United Kingdom
piunov@liv.ac.uk

Abstract: This article concerns the optimal stopping problem for a discrete-time Markov chain with observable states, but with unknown transition probabilities. A stopping policy is graded via the expected total-cost criterion resulting from the non-negative running and terminal costs. The Dynamic Programming method, combined with the Bayesian approach, is developed. A series of explicitly solved meaningful examples illustrates all the theoretical issues.

Keywords: Markov Decision Process (MDP), unknown transition matrices, dynamic programming, Bayesian method, optimal stopping

1. Introduction

In this article, we consider a statistical problem in which we observe the state of the system completely at each step, but the stochastic law of the system is unknown. The objective is to stop the process in such a possible way that the total expected running costs and terminal cost are minimal. We formulate this model as a Markov decision process (MDP) with unknown transition matrix.

The learning algorithm and the adaptive control problem in MDP have been studied by many authors: Easley and Kiefer (1988), Wang and Yi (2009), van Hee (1978), Hernández-Lerma and Marcus (1985), Hernández-Lerma (1989), Kurano (1972, 1983), Mandl (1974), Martin (1967).

For a general discussion of the Bayesian dynamic decision model, see van Hee (1978), Martin (1967), Rieder (1975).

[†]Submitted: January 2013; Accepted: July 2013

Adaptive control processes, such as the statistical design of sequential sampling problem, may be considered as stopping problems in MDP, White (1969); the Dynamic Programming (DP) approach proved to be effective here. Another Bayesian sequential analysis based on the DP approach has been presented in Ross (1970, 1983).

The statistical decision function approach was studied in DeGroot (1970), Raiffa and Schlaifer (1961), Wald (1950). The optimization problem in MDP under the total expected cost criterion can be also considered as a stopping problem: Hordijk (1974).

In Piunovskiy (2006), constrained discounted MDP was studied using the Dynamic Programming approach. Recently, constrained optimal stopping problems were considered using the convex analytic approach in Dufour and Piunovskiy (2010), Horiguchi (2001a,b). The optimal stopping theory is used for solving selection problems in Stadje (1997), where offers are successively available and the decision maker has to select one of them. In the article mentioned, a special case is studied in the framework of incomplete information. A more specific application, namely, selling an asset, was intensively discussed in the last years: see Section 10.3.1 in Bäuerle and Rieder (2011), and in Ekström and Lu (2011) in the continuous-time framework. Among the recent works on partially observed MDPs, let us mention Easley and Kiefer (1988) and Wang and Yi (2009). In both these articles, the total discounted expected reward was studied and, in the latter case, a specific model for the unknown parameter was presented. Application to the optimal investment-consumption along with the portfolio optimization is presented in Wang and Yi (2009).

In the current paper, we treat the statistical problem as a stopping problem in MDP using the Bayesian method and the DP approach. For a special, but quite general case, we prove that the optimal stopping rule is of threshold type. A big part of the current paper is devoted to meaningful examples, like preventive maintenance of a production line. In Section 2, we describe the general problem under study, and in Section 3 we develop the DP approach combined with the Bayesian method. Section 4 is for numerous meaningful explicitly solved examples.

To the best of our knowledge, the pure undiscounted stopping problem with partial information was not considered earlier. Although the Bayesian approach is well known, this specific problem leads to a relatively simple Bellman equation which can be explicitly solved in several specific meaningful situations. Except for the novelty of the considered mathematical model, the current article contains a series of real-life applications (Section 4) useful for specialists in production management and for investors working in a random economic environment.

2. Problem statement

We consider the discrete-time Markov Decision Process (MDP) with the finite state space $S = \{1, 2, \dots, M\}$ and the uncontrolled transition matrix $Q^\eta =$

$[Q_{ij}^{\eta}]_{i,j\in S}$ measurably depending on the unknown paramter $\eta \in \Xi$, where $(\Xi, \mathcal{B}(\Xi))$ is a Borel space. If the current state is $i \in S$, the one-step cost equals $c(i)$. At any moment, the process can be stopped (sent to the absorbing state Δ without any future cost) with the terminal cost $C(i)$. The goal is to minimize the total expected cost.

This problem can be reformulated as a standard MDP with total cost in the following way. Let Δ denote the artificial state meaning the process is stopped, so that the state space becomes $S \cup \{\Delta\}$. The action space is $A = \{0, 1\}$, where $a = 0(1)$ means 'do not stop' ('stop the process'). Now the transition probability $\hat{Q}^{\eta}$ and the one-step cost $\hat{c}$ in the state $i \in S \cup \{\Delta\}$ are defined as

$$\hat{Q}^{\eta}(j|i,a) = \begin{cases} Q_{ij}^{\eta}, & \text{if } i,j \neq \Delta,\ a = 0; \\ 1, & \text{if } j = \Delta \text{ and } i = \Delta \text{ or } a = 1; \\ 0 & \text{otherwise.} \end{cases}$$

$$\hat{c}(i,a) = \begin{cases} c(i), & \text{if } i \neq \Delta \text{ and } a = 0; \\ C(i), & \text{if } i \neq \Delta \text{ and } a = 1; \\ 0, & \text{if } i = \Delta. \end{cases}$$

We assume that the initial probability distribution ν on $S \cup \{\Delta\}$ of the state x_0 is given, as well as the initial distribution $F^0(d\eta)$ of parameter η. Below, $x_0, x_1, x_2, \ldots$ is the observable trajectory of the controlled process in the space $S \cup \{\Delta\}$. On each step $t = 0, 1, 2, \ldots$, the decision $a_t \in A$ must be chosen.

Let Π be the set of all randomized past dependent control policies $\pi = \{\pi_t\}_{t=0,1,2,\ldots}$ where, π_t is a stochastic kernel on the action space A given $(S \cup \{\Delta\}) \times \mathcal{P}(\Xi) \times (A \times (S \cup \{\Delta\}))^t$. Here, $\mathcal{P}(\Xi)$ is the Borel space of all probability measures on $(\Xi, \mathcal{B}(\Xi))$ (see Bertsekas and Shreve, 1978). Define $\Omega = (S \cup \{\Delta\}) \times \Xi \times (A \times (S \cup \{\Delta\}))^{\infty}$ and let $\mathcal{F}$ be its associated product σ-algebra. Similarly to Section 2.2 in Hernández-Lerma and Lasserre (1996), for an arbitrary policy $\pi \in \Pi$, there exists a probability measure P_{ν,F^0}^{π} on $(\Omega, \mathcal{F})$ such that the unknown parameter η and the coordinate projections x_t (respectively a_t) from Ω to the set $S \cup \{\Delta\}$ (respectively A) satisfy

(i) $P_{\nu,F^0}^{\pi}[\eta \in B] = F^0(B)$;
(ii) $P_{\nu,F^0}^{\pi}[x_0 = i] = \nu(i)$;
(iii) $P_{\nu,F^0}^{\pi}[a_t = l|\mathcal{F}_t] = \pi_t(l|f_t)$;
(iv) $P_{\nu,F^0}^{\pi}[x_{t+1} = j|\mathcal{F}_t \vee \sigma\{a_t\}] = \hat{Q}^{\eta}(j|x_t, a_t)$,

for any $B \in \mathcal{B}(\Xi)$, $i, j \in S \cup \{\Delta\}$, $l \in A$, where $\mathcal{F}_t = \sigma(f_t)$ with $f_0 = (x_0, F^0)$, $f_t = (x_0, F^0, a_0, x_1, a_1, \ldots, x_t)$.

For more about control policies and strategic measures the Reader is referred to Bertsekas and Shreve (1978), Dynkin and Yushkevich (1979), Hernández-Lerma and Lasserre (1996).

The target is to minimize the total expected cost

$$E_{\nu,F^0}^{\pi}\left[\sum_{t=0}^{\infty} \hat{c}(x_t, a_t)\right] \longrightarrow \inf_{\pi}.$$

Here and below, E^{π}_{ν,F^0} is the mathematical expectation with respect to the measure P^{π}_{ν,F^0}.

Another, equivalent description of the optimal stopping problem can be found in the article by Dufour and Piunovskiy (2010), devoted to the case of the complete information.

Note that all the results presented in the current paper obviously hold in case the state space S is countable.

3. Dynamic programming approach

It is known that the *a posteriori* probability distributions

$$F^t(\Gamma) = P^{\pi}_{\nu,F^0}(\eta \in \Gamma | \mathcal{F}_t), \quad \Gamma \in \mathcal{B}(\Xi), \qquad t = 0, 1, 2, \ldots,$$

along with the observed states x_t, form the sufficient statistic: Dynkin and Yushkevich (1979), Section 3.3 in Ferguson (1967). Such distributions can be recursively calculated using the Bayes formula, up to the stopping moment. If the distribution $F^{t-1}(d\eta)$ is known and transition $i = x_{t-1} \to x_t = j$ was observed, with $i, j \neq \Delta$, then

$$F^t(\Gamma) = \frac{\int_{\Gamma} Q^{\eta}_{ij} F^{t-1}(d\eta)}{\int_{\Xi} Q^{\eta}_{ij} F^{t-1}(d\eta)}. \tag{1}$$

(Compare with the formula on p.1048 in Easley and Kiefer, 1988.) Here and below, the case $\int_{\Xi} Q^{\eta}_{ij} F^{t-1}(d\eta) = 0$ is excluded because otherwise the transition $i \to j$ has zero probability. Therefore, all the expressions in the denominators are positive. Now, the pair (x_t, F^t) is a fully observed random process. For the given current values $(x_{t-1} = i, F^{t-1})$, the probability of the new state $x_t = j$ equals $\int_{\Xi} Q^{\eta}_{ij} F^{t-1}(d\eta)$ and, if $x_{t-1} = i$ and $x_t = j$, then the new component F^t is non-random, calculated using equation (1). The above calculations are valid for $i, j \neq \Delta$, $a_{t-1} = 0$. Action $a_{t-1} = 1$ (stop the process) at time moment $t-1$ results in the absorption at state (Δ, F^{t-1}). The cost function does not depend on the component F^{t-1} and coincides with the function $\hat{c}$. Below, we investigate the constructed MDP with complete information.

Assumption 1 *All costs are non-negative:* $c, C \geq 0$.

The dynamic programming approach for MDP satisfying Assumption 1 is well known. For $i \in S$, the Bellman equation looks as follows:

$$V(i, F) = \min \left\{ c(i) + \int_{\Xi} \sum_{j \in S} Q^{\eta}_{ij} F(d\eta) V(j, G_{ij} \circ F); \quad C(i) \right\}. \tag{2}$$

Here, G_{ij} is the operator on the right-hand side of equation (1): for $F \in \mathcal{P}(\Xi)$, $G_{ij} \circ F$ is the probability measure on Ξ defined by

$$G_{ij} \circ F(\Gamma) = \frac{\int_\Gamma Q_{ij}^\eta F(d\eta)}{\int_\Xi Q_{ij}^\eta F(d\eta)}, \quad \Gamma \in \mathcal{B}(\Xi).$$

Obviously, $V(\Delta, F) \equiv 0$.

Before the process is stopped, the Bellman function $V(i, F)$ depends on the current state i of the process x_t and the current *a posteriori* distribution F of parameter η. Equation (2) can be solved by successive approximations

$$V_0(i, F) = 0, \qquad V_{n+1}(i, F) = H \circ V_n(i, F),$$

where H is the Bellman operator presented on the right-hand side of formula (2). According to Corollary 9.17.1 in Bertsekas and Shreve (1978), the Bellman function V is the minimal non-negative solution of equation (2); see also Section 7.2.8 in Puterman (1994).

Instead of the distributions F^t of the parameter η, one can consider the images $\tilde{F}^t$ with respect to the mapping $\eta \to Q$, probability distributions of stochastic matrices Q. For fixed $i, j \in S$, let $\tilde{F}^t_{(ij)}(dQ_{ij})$ be the marginal distribution of the component $Q_{ij} \in [0, 1]$ and let $\tilde{F}^t_{\cdot|(ij)}(\cdot|Q_{ij})$ be the conditional distribution of all other elements of the Q matrix under a given value Q_{ij}. Then, according to (1),

$$\tilde{F}^t_{(ij)}(\Gamma^{(ij)}) = \int_\Xi 1\{Q_{ij}^\eta \in \Gamma^{(ij)}\} F^t(d\eta) = \frac{\int_\Xi 1\{Q_{ij}^\eta \in \Gamma^{(ij)}\} Q_{ij}^\eta F^{t-1}(d\eta)}{\int_\Xi Q_{ij}^\eta F^{t-1}(d\eta)},$$

and the change of variable $z = Q_{ij}^\eta$ implies

$$\tilde{F}^t_{(ij)}(\Gamma^{(ij)}) = \frac{\int_{\Gamma^{(ij)}} z \cdot \tilde{F}^{t-1}_{(ij)}(dz)}{\int_{[0,1]} z \cdot \tilde{F}^{t-1}_{(ij)}(dz)}, \qquad \Gamma^{(ij)} \in \mathcal{B}([0, 1]).$$

Further, for any $\Gamma \in \mathcal{B}([0, 1]^{M^2})$,

$$\tilde{F}^t(\Gamma) = \int_\Xi 1\{Q^\eta \in \Gamma\} F^t(d\eta) = \frac{\int_\Xi 1\{Q^\eta \in \Gamma\} Q_{ij}^\eta F^{t-1}(d\eta)}{\int_\Xi Q_{ij}^\eta F^{t-1}(d\eta)}.$$

In the numenator, we introduce the variables $z = Q_{ij}^\eta \in [0, 1]$ and $y \in [0, 1]^{M^2-1}$; the latter denotes all the remainder components of the Q^η matrix apart from

Q^{η}_{ij}. Now, using this change of variables and the Fubini Theorem, we obtain

$$\begin{aligned}\tilde{F}^t(\Gamma) &= \frac{\displaystyle\int_{[0,1]}\int_{[0,1]^{M^2-1}} 1\{y\in\Gamma_z\}1\{z\in\Gamma^{(ij)}\}z\cdot\tilde{F}^{t-1}_{\cdot|(ij)}(dy|z)\tilde{F}^{t-1}_{(ij)}(dz)}{\displaystyle\int_{[0,1]} z\cdot\tilde{F}^{t-1}_{(ij)}(dz)}\\ &= \frac{\displaystyle\int_{\Gamma^{(ij)}}\left[z\cdot\tilde{F}^{t-1}_{\cdot|(ij)}(\Gamma_z|z)\right]\tilde{F}^{t-1}_{(ij)}(dz)}{\displaystyle\int_{[0,1]} z\cdot\tilde{F}^{t-1}_{(ij)}(dz)},\end{aligned}$$

where $\Gamma^{(ij)}=\{z\in\mathbb{R}:\ \exists Q\in\Gamma \text{ with } Q_{ij}=z\}$ and $\Gamma_z\subset\mathbb{R}^{M^2-1}$ is the z-section of the set Γ (all possible values of the components of the Q matrix, apart from the component $Q_{ij}=z$, such that $Q\in\Gamma$).

We see that

$$\tilde{F}^t(\Gamma)=\int_{\Gamma^{(ij)}}\tilde{F}^{t-1}_{\cdot|(ij)}(\Gamma_z|z)\cdot\tilde{F}^{t}_{(ij)}(dz)$$

meaning that

$$\tilde{F}^t_{\cdot|(ij)}(dy|z)=\tilde{F}^{t-1}_{\cdot|(ij)}(dy|z):$$

the conditional distribution $\tilde{F}_{\cdot|(ij)}$ does not change if one observes the transition $i\to j$ of the x_t process.

Let us mention briefly several possible generalizations of the model for which the presented dynamic programming approach can be developed, too.

Firstly, the one-step cost $c(i,\eta)$ and the terminal cost $C(i,\eta)$ can depend on the unknown parameter η. In this case one should modify the cost $\hat{c}$:

$$\hat{c}(i,a,F)=\begin{cases}\displaystyle\int_{\Xi}c(i,\eta)F(d\eta), & \text{if } i\neq\Delta \text{ and } a=0;\\ \displaystyle\int_{\Xi}C(i,\eta)F(d\eta), & \text{if } i\neq\Delta \text{ and } a=1;\\ 0, & \text{if } i=\Delta.\end{cases}$$

Secondly, the state space S may be arbitrary Borel and the dynamics of the process is given by the measurable stochastic kernel $Q^{\eta}(dy|x)$.

In these situations, the operator for the *posterior* distribution (1) must be modified in the following way. If the distribution $F^{t-1}(d\eta)$ is known and the current state is $x_{t-1}=x\in S$, then the joint distribution of η, the cost c, and the next state y is given by

$$P_x(\Gamma^{\eta}\times\Gamma^{c}\times\Gamma^{y})=\int_{\Gamma^{\eta}}Q^{\eta}(\Gamma^{y}|x)1\{c(x,\eta)\in\Gamma^{c}\}F^{t-1}(d\eta),$$

where $\Gamma^\eta \in \mathcal{B}(\Xi), \Gamma^c \in \mathcal{B}(\mathbb{R}), \Gamma^y \in \mathcal{B}(S)$ are arbitrary Borel subsets. This probability measure can be decomposed as

$$P_x(\Gamma^\eta \times \Gamma^c \times \Gamma^y) = \int_{\Gamma^c \times \Gamma^y} P_x^m(dc \times dy)\Phi_x(\Gamma^\eta | c, y),$$

where $P_x^m(\Gamma^c \times \Gamma^y) = P_x(\Xi \times \Gamma^c \times \Gamma^y)$ is the marginal distribution of the elements (c, y), and Φ_x is a measurable stochastic kernel on Ξ given $\mathbb{R} \times S$. (See Bertsekas and Shreve, 1978.) Now $F^t(\Gamma^\eta) = \Phi_x(\Gamma^\eta | c, y)$ is the *a posteriori* distribution of the parameter η after the cost $c = c(x, \eta)$ and the next state of the process $y = x_t$ are observed. One should replace the operator G_{ij} with G_{xyc}:

$$F^t(\cdot) = G_{xyc} \circ F^{t-1}(\cdot) = \Phi_x(\cdot | c, y).$$

The Bellman equation (2) takes the form

$$\begin{aligned} V(x, F) &= \min \left\{ \int_\Xi \left[c(x, \eta) + \int_S Q^\eta(dy|x)\ V(y, G_{xyc(x,\eta)} \circ F) \right] F(d\eta); \right. \\ &\qquad \left. \int_\Xi C(x, \eta) F(d\eta) \right\}. \end{aligned}$$

4. Examples

Consider a game where the player, after paying the entrance fee $c > 0$, can win with probability p and lose with probability $q = 1 - p$. After winning, the reward equals $R > 0$ and the player will not play any more. After losing, the player can try again after paying the same fee c, or stop the process with the terminal cost $C \geq 0$. Probability $q = \eta$ is unknown with a given initial distribution $F^0(d\eta)$ on $\Xi = [0, 1]$. Therefore, we deal with the state space $S = \{0, 1\}$, where 0(1) means the player has lost (has won) the previous game; $Q_{00} = q$, $Q_{11} = 1$. If we add the constant R to the total expected cost, we can replace the terminal rewards with regrets: $C(1) = 0$, $C(0) = C + R$, so that Assumption 1 will be satisfied. The one-step costs are equal $c(0) = c$, $c(1) = 0$. Below, we assume that $\int_{[0,1]} \eta F^0(d\eta) \in (0, 1)$.

Since the model is positive, the Bellman function $V(i, F)$ is non-negative, and, from equation (2), it immediately follows that

$$\begin{aligned} V(1, F) &= \min\{0 + V(1, F); \quad 0\} = 0; \\ V(0, F) &= \min \left\{ c(0) + \int_0^1 (1 - \eta) dF(\eta) V(1, G_{01} \circ F) \right. \\ &\qquad \left. + \int_0^1 \eta dF(\eta) V(0, G_{00} \circ F); \quad C(0) \right\}. \end{aligned}$$

Here, with some abuse of notation, we replace the distribution F with $F(\eta)$, the cumulative distribution function (CDF) of η on the interval $[0, 1]$.

Since $V(1,F)=0$ does not depend on F, we rewrite the main equation:

$$V(0,F)=\min\left\{c(0)+\int_0^1 \eta dF(\eta)V(0,G_{00}\circ F);\quad C(0)\right\}, \tag{3}$$

where, according to (1),

$$G_{00}\circ F(\eta)=\frac{\int_0^\eta ydF(y)}{\int_0^1 ydF(y)}. \tag{4}$$

If the first (second) expression in (3) is smaller then the optimal action in state 0, having the *aposteriori* distribution F, is $a=0$ ($a=1$).

Let $F^n=G_{00}\circ F^{n-1}$ with the given initial CDF F^0 on the interval $[0,1]$. Let

$$q_n=\int_0^1 \eta dF^n(\eta)$$

and let

$$q_\infty=\min\{y:\ F^0(y)=1\}.$$

Lemma 1 *(a)*

$$F^n(\eta)=\frac{\int_0^\eta y^n dF^0(y)}{q_0q_1\dots q_{n-1}}=\frac{\int_0^\eta y^n dF^0(y)}{\int_0^1 y^n dF^0(y)}; \tag{5}$$

$$q_0q_1\dots q_n=\int_0^1 y^{n+1}dF^0(y). \tag{6}$$

(b) If $\eta\ge q_\infty$ then $\lim_{n\to\infty}F^n(\eta)=1$; if $\eta<q_\infty$ then $\lim_{n\to\infty}F^n(\eta)=0$.

(c) The sequence q_n is non-decreasing: $\forall n=0,1,2,\dots$, $q_{n+1}\ge q_n$, and $\lim_{n\to\infty}q_n=q_\infty$.

Proof. (a) The left equality (5) is valid for $n=0$. (Here $q_0q_1\dots q_{n-1}=1$.) If it holds for some $n\ge 0$ then

$$F^{n+1}(\eta)=\frac{\int_0^\eta ydF^n(y)}{q_n}=\frac{\int_0^\eta y^{n+1}dF^0(y)}{q_0q_1\dots q_n}.$$

If we substitute $\eta=1$, we obtain formula (6).

Now the right equality (5) is also obvious.

(b) According to (5), $\forall\eta\ge q_\infty$, $\forall n=0,1,2,\dots$, $F^n(\eta)\equiv 1$.

In case $F^0(\eta)=0$, $F^n(\eta)\equiv 0$.

Suppose now that $\eta < q_\infty$ and $F^0(\eta) > 0$. Let $\tilde{\eta} \in (\eta, q_\infty)$, so that $F^0(\tilde{\eta}) < 1$. Since, for any $n = 0, 1, 2, \ldots$, $F^n(\eta) > 0$, we can consider

$$R^n = \frac{1}{F^n(\eta)} = 1 + \frac{\int_\eta^1 y^n dF^0(y)}{\int_0^\eta y^n dF^0(y)}.$$

(See formula (5).)

Now

$$R^n \geq \frac{\int_{\tilde{\eta}}^1 y^n dF^0(y)}{\int_0^\eta y^n dF^0(y)} \geq \frac{\tilde{\eta}^n [1 - F^0(\tilde{\eta})]}{\eta^n F^0(\eta)} = \left(\frac{\tilde{\eta}}{\eta}\right)^n \frac{[1 - F^0(\tilde{\eta})]}{F^0(\eta)}$$

and $\lim_{n\to\infty} R^n = \infty$ meaning that $\lim_{n\to\infty} F^n(\eta) = 0$.

(c) According to (4),

$$q_{n+1} = \frac{\int_0^1 \eta^2 dF^n(\eta)}{\int_0^1 \eta dF^n(\eta)}.$$

Therefore,

$$q_{n+1} - q_n = \frac{1}{\int_0^1 \eta dF^n(\eta)} \left[\int_0^1 \eta^2 dF^n(\eta) - \left(\int_0^1 \eta dF^n(\eta)\right)^2\right] \geq 0$$

and the limit $\lim_{n\to\infty} q_n$ exists.

For any $n = 0, 1, 2, \ldots$,

$$q_n = \int_0^{q_\infty} \eta dF^n(\eta) \leq q_\infty.$$

(Remember that $F^n(q_\infty) = 1$.)

Fix an arbitrary $\varepsilon > 0$. Now

$$q_n = \int_0^{q_\infty - \varepsilon} \eta dF^n(\eta) + \int_{q_\infty - \varepsilon}^{q_\infty} \eta dF^n(\eta)$$

$$\geq (q_\infty - \varepsilon)[F^n(q_\infty) - F^n(q_\infty - \varepsilon)].$$

Since $F^n(q_\infty) \equiv 1$ and $\lim_{n\to\infty} F^n(q_\infty - \varepsilon) = 0$, we conclude that

$$\lim_{n\to\infty} q_n \geq q_\infty - \varepsilon.$$

As $\varepsilon > 0$ was arbitrarily small, $\lim_{n\to\infty} q_n = q_\infty$.

The proof is completed.

Theorem 1 *(a) If*

$$c(0) \le C(0)(1 - q_\infty)$$

then the optimal action in state 0 is $a_t \equiv 0$ *(do not stop the process) and, for all distributions* F^0, $F^1 = G_{00} \circ F^0$, ..., $F^n = G_{00} \circ F^{n-1}$, ...,

$$V(0, F^n) = \frac{c(0) \int_0^1 \frac{\eta^n}{1-\eta} dF^0(\eta)}{\int_0^1 \eta^n dF^0(\eta)}.$$

(Other distributions are never realized.)

(b) In case

$$c(0) > C(0)(1 - q_\infty)$$

let $n^* \ge 0$ *be the first integer such that* $c(0) > C(0)(1 - q_n)$.

Then action $a_t = 0$ *is optimal in state 0 after the first* n^* *observations of that state. If the state 0 is observed* $(n^* + 1)$ *times, then it is optimal to stop the process immediately.*

If $n \ge n^*$, *then* $V(0, F^n) = C(0)$; *if* $n < n^*$, *then*

$$V(0, F^n) = c(0) \sum_{i=1}^{n^*-n} \prod_{k=0}^{i-2} q_{n+k} + C(0) \prod_{k=0}^{n^*-n-1} q_{n+k}.$$

Here, as usual, $\sum_{j=1}^{0} z_j = 0$ *and* $\prod_{k=1}^{0} z_k = 1$.

Proof. (a) We solve equation (3) by successive approximations:

$$\begin{aligned} V_0(0, F^n) &\equiv 0, \\ V_{i+1}(0, F^n) &= \min\{c(0) + q_n V_i(0, F^{n+1}); \ C(0)\}. \end{aligned} \tag{7}$$

This solution is given by the formula

$$V_i(0, F^n) = c(0) \sum_{j=1}^{i} \prod_{k=1}^{j-1} q_{n+k-1}. \tag{8}$$

The presented formula is correct for $i = 0$. If it holds for some $i \geq 0$ then

$$\begin{aligned} c(0) + q_n V_i(0, F^{n+1}) &= c(0) + q_n c(0) \sum_{j=1}^{i} \prod_{k=1}^{j-1} q_{n+k} \\ &= c(0) \left[1 + \sum_{j=1}^{i} \prod_{k=0}^{j-1} q_{n+k} \right] \\ &= c(0) \left[1 + \sum_{j=2}^{i+1} \prod_{k=1}^{j-1} q_{n+k-1} \right] \\ &= c(0) \sum_{j=1}^{i+1} \prod_{k=1}^{j-1} q_{n+k-1} \end{aligned}$$

and, for any $i = 0, 1, 2, \ldots$, for all $n \geq 0$

$$c(0) \sum_{j=1}^{i} \prod_{k=1}^{j-1} q_{n+k-1} \leq c(0) \sum_{j=1}^{\infty} q_\infty^{j-1} = \frac{c(0)}{1 - q_\infty} \leq C(0).$$

Let us pass to the limit as $i \to \infty$ in the formula (8). (Remember that $q_n \leq q_\infty < 1$.)

$$\begin{aligned} V(0, F^n) &= \lim_{i \to \infty} V_i(0, F^n) = c(0) \sum_{j=1}^{\infty} \prod_{k=1}^{j-1} q_{n+k-1} \\ &= \frac{c(0) \sum_{j=1}^{\infty} \prod_{k=0}^{n+j-2} q_k}{q_0 q_1 \cdots q_{n-1}} = \frac{c(0) \sum_{j=1}^{\infty} \int_0^1 \eta^{n+j-1} dF^0(\eta)}{\int_0^1 \eta^n dF^0(\eta)}. \end{aligned}$$

Here we used the obvious expression

$$\prod_{k=0}^{n+j-2} q_k = \prod_{k=0}^{n-1} q_k \prod_{k=1}^{j-1} q_{n+k-1}$$

and formula (6). Finally, by the Lebesgue monotone convergence theorem,

$$V(0, F^n) = \frac{c(0) \int_0^1 \frac{\eta^n}{1 - \eta} dF^0(\eta)}{\int_0^1 \eta^n dF^0(\eta)}.$$

(b) Firstly, let us prove that, for all $n \geq n^*$, $V(0, F^n) = C(0)$, so that, if the state 0 is observed $(n^* + 1)$ times, then it is optimal to stop the process. (Under the optimal policy, the state 0 cannot be observed more times.)

Suppose that there is N such that for all $n \geq N$

$$V(0, F^n) = c(0) + q_n V(0, F^{n+1}) < C(0).$$

We know that the sequence q_n increases. Since $V_0(0, F^n) \equiv 0$ is not decreasing with n, the sequence $\{V_i(0, F^n)\}_{n=N}^{\infty}$ is also non-decreasing for any $i = 1, 2, \ldots$, so the limiting sequence $\{V(0, F^n)\}_{n=N}^{\infty}$ is also non-decreasing and bounded. Thus, there is a limit $W = \lim_{n\to\infty} V(0, F^n)$ satisfying equation

$$W = c(0) + q_\infty W.$$

On the one hand, $W \leq C(0)$, but on the other hand

$$W = \frac{c(0)}{1 - q_\infty} > C(0).$$

The obtained contradiction shows that, for any n, there is $K > n$ such that $V(0, F^K) = C(0)$.

Let $n \geq n^*$ be fixed and take the corresponding integer $K > n$. Now, considering sequentially $i = K - 1, K - 2, \ldots, n$, we see that

$$c(0) + q_i V(0, F^{i+1}) = c(0) + q_i C(0)$$

$$> C(0)(1 - q_i) + q_i C(0) = C(0).$$

(Remember that $c(0) > C(0)(1 - q_i)$.) Therefore, $V(0, F^n) = C(0)$.

For $n < n^*$, let us prove by induction that the presented expression for $V(0, F^n)$ satisfies the Bellman equation and is smaller than $C(0)$.

For $n = n^* - 1$, we have

$$\begin{aligned} V(0, F^n) &= c(0) + q_n C(0) = c(0) + q_n V(0, F^{n+1}) \\ &\leq C(0)(1 - q_n) + q_n C(0) = C(0). \end{aligned}$$

Suppose the Bellman equation is satisfied at some $n \leq n^* - 1$, $V(0, F^n) \leq C(0)$, and consider $n - 1$:

$$\begin{aligned} & c(0) + q_{n-1} V(0, F^n) \leq c(0) + q_{n-1} C(0) \\ \leq\ & C(0)(1 - q_{n-1}) + q_{n-1} C(0) = C(0); \end{aligned}$$

$$\begin{aligned} c(0) + q_{n-1} V(0, F^n) &= c(0) + q_{n-1} \left\{ c(0) \sum_{i=1}^{n^*-n} \prod_{k=0}^{i-2} q_{n+k} \right. \\ &\qquad \left. + C(0) \prod_{k=0}^{n^*-n-1} q_{n+k} \right\} \\ &= c(0) \sum_{i=1}^{n^*-n+1} \prod_{k=0}^{i-2} q_{n-1+k} + C(0) \prod_{k=0}^{n^*-n} q_{n-1+k}. \end{aligned}$$

The proof is completed.

REMARK 1 *If we consider maximization of the total expected reward, minimum in equation (3) should be replaced with maximum. Again, we consider the positive model with* $c(0) > 0,\ C(0) \geq 0$. *In this case, the analogue of Theorem 1 is simpler and looks as follows. If* $c(0)\int_0^1 \frac{1}{1-\eta} dF^0(\eta) > C(0)$ *then never stop the process; otherwise, stop immediately. For the proof, note that the case*

$$V(0, F^n) = c(0) + q_n V(0, F^{n+1}) \text{ and } V(0, F^{n+1}) = C(0)$$

(for some $n \geq 0$*) is excluded because*

$$c(0) + q_{n+1} V(0, F^{n+2}) > c(0) + q_n C(0) = V(0, F^n) \geq C(0).$$

(We omit the trivial situation when $q_n = q_{n+1}$, *i.e. the distribution* F^n *is degenerate.) Now one should consider only two control policies: never stop the process, or stop immediately, for which the total expected rewards equal* $c(0)\int_0^1 \frac{1}{1-\eta} dF^0(\eta)$ *and* $C(0)$, *correspondingly. The Bellman function is given by*

$$V(0, F) = \begin{cases} c(0)\int_0^1 \frac{1}{1-\eta} dF(\eta), & \text{if } c(0)\int_0^1 \frac{1}{1-\eta} dF(\eta) > C(0); \\ C(0) & \text{otherwise.} \end{cases}$$

The case $V(0, F) = \infty$ *is not excluded here.*

If there is no information about the unknown probability $\eta = q$ then it is standard practice to consider the uniform distribution

$$F^0(\eta) = \begin{cases} 0, & \text{if } \eta < 0, \\ \eta, & \text{if } \eta \in [0, 1], \\ 1, & \text{if } \eta > 1. \end{cases}$$

In this case

$$q_\infty = 1, \qquad q_n = \frac{n+1}{n+2},$$

so that we deal with the part (b) of Theorem 1. Let $n^* \geq 0$ be the first integer such that $(n+2)c(0) > C(0)$. Then, when in state 0, the process must be stopped after observing state 0 (n^*+1) times. The Bellman function is given by the following formulae:

$$\begin{aligned} V(1, F) &\equiv 0, \\ V(0, F^n) &= C(0) \text{ for } n \geq n^*, \\ V(0, F^n) &= c(0)\left[(n+1)\left(\frac{1}{n+1} + \frac{1}{n+2} + \ldots + \frac{1}{n^*}\right)\right] \\ &\quad + \frac{n+1}{n^*+1} C(0), \quad \text{if } n < n^*. \end{aligned}$$

CDF $F^n(\eta)$ appears after the $(n+1)$-st observation of state 0; $n = 0, 1, 2, \ldots$; it has the form

$$F^n(\eta) = \begin{cases} 0, & \text{if } \eta < 0, \\ \eta^{n+1}, & \text{if } \eta \in [0,1], \\ 1, & \text{if } \eta > 1. \end{cases}$$

Another special case corresponds to the beta-distribution:

$$f^0(\eta) = \frac{dF^0(\eta)}{d\eta} = \begin{cases} \frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \eta^{\nu_1-1}(1-\eta)^{\nu_2-1}, & \text{if } \eta \in [0,1], \\ 0 & \text{otherwise,} \end{cases}$$

where $\nu_1, \nu_2 > 0$ are constants.

In this case

$$q_\infty = 1, \qquad q_n = \frac{\nu_1 + n}{\nu_1 + \nu_2 + n},$$

so that we again deal with the part (b) of Theorem 1. Let $n^* \geq 0$ be the first integer such that $\left(\frac{\nu_1+\nu_2+n}{\nu_2}\right) c(0) > C(0)$. Then, when in state 0, the process must be stopped after observing state 0 (n^*+1) times. CDF $F^n(\eta)$ appears after the $(n+1)$-st observation of state 0; $n = 0, 1, 2, \ldots$; it is differentiable and follows the beta distribution:

$$f^n(\eta) = \frac{dF^n(\eta)}{d\eta} = \begin{cases} \frac{\Gamma(\nu_1+\nu_2+n)}{\Gamma(\nu_1+n)\Gamma(\nu_2)} \eta^{\nu_1+n-1}(1-\eta)^{\nu_2-1}, & \text{if } \eta \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

As always, $V(1, F) \equiv 0$, $V(0, F^n) = C(0)$ for $n \geq n^*$. For $n < n^*$, the expression for $V(0, F^n)$, presented in Theorem 1(b), cannot be essentially simplified, unless $\nu_2 = 1$ when

$$V(0, F^n) = c(0) \left[(n+\nu_1) \left(\frac{1}{n+\nu_1} + \frac{1}{n+\nu_1+1} + \ldots \right.\right.$$

$$\left.\left. + \frac{1}{n^*+\nu_1-1} \right) \right] + \frac{n+\nu_1}{n^*+\nu_1} C(0).$$

It is interesting to look at the case of $\nu_1 = \nu_2 = \nu$.

If $2c(0) > C(0)$ then, for any $\nu > 0$, we have $n^* = 0$ and $V(0, F^n) = C(0)$ for all $n \geq 0$.

If $2c(0) = C(0)$ then, for any $\nu > 0$, we have $n^* = 1$ and $V(0, F^n) = C(0)$ for all $n \geq 1$; $V(0, F^0) = c(0) + \frac{1}{2} C(0)$.

If $2c(0) < C(0)$ then, for small ν (when $\nu \to 0$), we again have $n^* = 1$. In case $\nu \to \infty$, obviously $n^* \to \infty$. By the way, in the last case the beta-density converges to the Dirac measure at point $\frac{1}{2}$ for which $q_\infty = \frac{1}{2}$ and, according to Theorem 1(a), in this limiting case, it is optimal not to stop the process in state 0 $(n^* = \infty)$.

Remember that the uniform distribution is the special case of the beta-distribution when $\nu_1 = \nu_2 = 1$.

If there is no information about the unknown probability $\eta = q$, one can also use the game-theoretic approach to find admissible actions in state 0.

Let $W(\pi, \eta)$ be the total expected loss if the initial state is 0, the control policy π is applied, and $q = \eta$.

For the pessimistic scenario,

$$\min_{\pi} \max_{\eta} W(\pi, \eta) = C(0)$$

and the optimal minmax decision is to stop the process because otherwise $\max_{\eta} W(\pi, \eta)$ is provided by $\eta = 1$ (or by the maximal possible probability): the process remains in state 0 with the positive running cost $c(0)$.

For the optimistic scenario,

$$\min_{\pi} \min_{\eta} W(\pi, \eta) = \begin{cases} c(0), & \text{if } c(0) < C(0), \\ C(0) & \text{otherwise,} \end{cases}$$

and the optimal decision is to stop the process if $c(0) \geq C(0)$. If the process is not stopped in state 0, then $\min_{\eta} W(\pi, \eta)$ is provided by $\eta = 0$ leading to the final cost $c(0)$.

For more about the game-theoretic approach see González–Trejo, Hernández–Lerma and Hoyos–Reyes (2003).

The presented results can be useful on different occasions listed below.

Special cases. (a) Consider a production system that can be stopped for reconstruction. The system can be in one of three possible states $S = \{1, 2, 3\}$, where 1 is the normal state, 2 is the warning state, and 3 means the failure. The one-step costs $c(1) = 0 < c(2) < c(3)$ are known; the state of the production system is observable. The rewards coming from the reconstruction, i.e. the terminal rewards $R(1) > R(2) > R(3) \geq 0$, are known.

The cost $c(2)$ is associated with the malfunctioning of the system in the warning state. We do not consider the possibility to repair the system in this state because this can be too expensive. Hence transition $2 \to 1$ is excluded. On the opposite, the cost $c(3)$ includes the price to repair the system, and the next state will be 1. This price is acceptable (or even equals zero), e.g. if we consider the warranty period when the repairman arrives if the failure occurs, but not in the warning state.

Therefore, the transition probabilities $Q_{21} = 0$ and $Q_{31} = 1$ are known; other transition probabilities $(Q^{\eta}_{12}, Q^{\eta}_{13}, Q^{\eta}_{22}) = \eta$ form the unknown parameter η with a given initial distribution $F^0(d\eta)$ on Ξ, where the space Ξ of the possible values of the vector parameter η is obvious:

$$\Xi = \{\eta \geq 0 \text{ (component-wise)} : \ Q^{\eta}_{12} + Q^{\eta}_{13} \leq 1, \ Q^{\eta}_{22} \leq 1\}.$$

If the Markov chain is never stopped, the total expected cost equals $+\infty$, so that action 1 ('stop the process') will be ultimately applied. If we add the constant $R(1)$ to the total expected cost, we can replace the terminal rewards with *regrets* $C(1) = 0$; $C(2) = R(1) - R(2)$; $C(3) = R(1) - R(3)$, so that Assumption 1 will be satisfied.

Since the model is positive, the Bellman function $V(i,F)$ is non-negative, and, from equation (2), it immediately follows that

$$\begin{aligned} V(1,F) &= \min\left\{ c(1) + \int_{\Xi} \sum_{j\in S} Q^{\eta}_{1j} F(d\eta) V(j, G_{1j}\circ F); \quad C(1) \right\} \\ &= C(1) = 0. \end{aligned}$$

Therefore, it is reasonable to stop the process in state 1, and the transition probabilities Q^{η}_{1j} are of no importance. Below, we omit them and consider $\eta = Q^{\eta}_{22}$. With some abuse of notation, we replace the distribution F with $F(\eta)$, the CDF of η on the interval $[0,1]$.

The Bellman equation (2) takes the form

$$\begin{aligned} V(2,F) &= \min\Big\{ c(2) + \int_0^1 (1-\eta) dF(\eta) V(3, G_{23}\circ F) \\ &\qquad + \int_0^1 \eta dF(\eta) V(2, G_{22}\circ F); \quad C(2) \Big\}; \\ V(3,F) &= \min\{c(3); \quad C(3)\}. \end{aligned}$$

Since $V(3,F)$ is known and does not depend on F, we omit this argument and concentrate on the main equation:

$$\begin{aligned} V(2,F) &= \min\Big\{ c(2) + V(3)[1 - \int_0^1 \eta dF(\eta)] \\ &\qquad + \int_0^1 \eta dF(\eta) V(2, G_{22}\circ F); \quad C(2) \Big\}, \end{aligned} \tag{9}$$

where, according to (1),

$$G_{22}\circ F(\eta) = \frac{\int_0^{\eta} y dF(y)}{\int_0^1 y dF(y)}.$$

The obtained equation (9), up to notations, coincides with equation (3). Indeed, it is sufficient to denote state 2 as 0, put $V(0,F) = V(2,F) - V(3)$ and replace $c(2)$ with just $c(0)$ and replace $C(2) - V(3)$ with $C(0)$. The optimal solution to this stopping problem follows now from Theorem 1. Here $C(0) = C(2) - V(3)$ may be negative, and we are looking for the minimal solution to equation (3), bigger than $-V(3)$. Theorem 1 remains valid.

(b) Again consider a production system which can be in one of the three possible states $S = \{1,2,3\}$, where 1 is the normal state, 2 is the warning state, and 3 means the failure. The one-step rewards $c(1) > c(2) > 0$ are known; the state of the production system is observable. In state 3, the system must be repaired, i.e. the process should be stopped with the terminal cost $C(3)$. In state

1, there is no possibility to stop the process. Only in state 2 the decision maker must decide whether to stop the process (repair the system) with the terminal cost $C(2) < C(3)$, or not. The transition probability $Q_{22}^{\eta} = \eta$ is unknown, with a given initial distribution $F^0(d\eta)$ on $\Xi = [0,1]$; $Q_{21} = 0$; $Q_{23}^{\eta} = 1 - \eta$; other transition probabilities are of no importance, but we assume that $Q_{11} < 1$.

The dynamic programming approach leads to the following equations

$$\begin{aligned} V(1,F) &= c(1) + Q_{11}V(1,F) + Q_{12}V(2,F) + Q_{13}V(3,F); \\ V(2,F) &= \max\left\{ c(2) + \int_0^1 (1-\eta)dF(\eta)V(3, G_{23}\circ F) \right. \\ &\qquad \left. + \int_0^1 \eta dF(\eta)V(2, G_{22}\circ F); \quad -C(2) \right\}; \\ V(3,F) &= -C(3). \end{aligned}$$

Here, like previously, F is the CDF of η and

$$G_{22}\circ F(\eta) = \frac{\int_0^{\eta} y dF(y)}{\int_0^1 y dF(y)}.$$

If the transition probabilities Q_{11}, Q_{12}, Q_{13} are unknown, one should replace them with their *aposteriori* estimates.

Clearly, only the second equation must be investigated. After we denote $\tilde{V}(2,F) = V(2,F) + C(3)$, it takes the form

$$\tilde{V}(2,F) = \max\left\{ c(2) + \int_0^1 \eta dF(\eta)\tilde{V}(2, G_{22}\circ F);\ C(3) - C(2) \right\}.$$

Since $c(2) > 0$, $C(3) > C(2)$, the model is positive, and we are looking for the minimal non-negative solution. If needed, function $V(1,F)$ can be calculated/estimated using the first equation:

$$V(1,F) = \frac{c(1) + Q_{12}\tilde{V}(2,F)}{1 - Q_{11}} - C(3).$$

Clearly, the obtained equation coincides with (3) if we denote state 2 as 0, replace $C(3) - C(2)$ with $C(0)$, and consider the maximization problem (see Remark 1).

The optimal action in state 0 (in the original state 2) is

$$\begin{cases} \text{never stop the process if} \\ \quad c(0)\displaystyle\int_0^1 \frac{1}{1-\eta} dF(\eta) > C(0) \\ \\ \quad \Longleftrightarrow c(2)\displaystyle\int_0^1 \frac{1}{1-\eta} dF(\eta) > [C(3) - C(2)]; \\ \\ \text{stop immediately otherwise.} \end{cases}$$

This result is understandable. If the total expected reward in the warning state 2, given the *a priori* distribution F, is smaller than $C(3) - C(2)$, the reward coming from the repair in the warning state, compared with the failure state, then stop the process. Otherwise, do not stop. If the process is not stopped in state 2 then there is no reason to stop it later because, with each transition $2 \to 2$, the situation improves: the expected *a posteriori* total reward in state 2 increases.

(c) Playing the game can mean starting a project in a random economical environment. The entrance fee equals the investment, the reward R is the revenue in case the project is successful. One can put $C = 0$ in this case. The probability p of the favorable environment is unknown, and the question is whether it is reasonable to restart the project if the previous attempt was unsuccessful.

(d) Finally, one can consider the standard hide-and-seek game, similar to that discussed in Chapter III Section 5 of Ross (1983). Suppose that on each step the object can occupy any of the locations $\{a, b\}$ with the same probability $1/2$, so that it does not matter which location to examine. If the decision maker looks at the proper location, the probability that the object is discovered equals α, unknown probability. Thus, in this case $p = \alpha/2$, the reward for discovering the object equals R, the entrance fee for each one round is c. The question is whether to continue this game or stop the process after several trials.

Another modification corresponds to the case when the object is moving between the possible locations $\{a, b\}$ according to a Markov chain with unknown transition probabilities $p_{aa}, p_{ab}, p_{ba}, p_{bb}$. Suppose the player is unable to search location b and the probability to discover the object in the location a is $\alpha = 1$. Thus, transition probabilities p_{aa} and p_{ab} are of no importance. We assume that the initial probability of the object to be in the location b equals p_{bb}. Now we deal with the game discussed above, where the unknown probability of losing one round equals $\eta = q = p_{bb}$.

Suppose now that on each step the object is in location a with probability p_a, that is, the transition probabilities introduced above equal $p_{aa} = p_{ba} = p_a$, $p_{bb} = p_{ab} = p_b = 1 - p_a$. Again assume the player is unable to search location b and the probability to discover the object in the location a is $\alpha \in [0, 1]$. Now we have the game discussed above, where the unknown probability of losing one round equals $\eta = q = p_b + p_a(1-\alpha) = 1 - \alpha p_a$. Both probabilities, α and p_a may be unknown, but the developed theory operates with the compound probability η.

In all these cases, Theorem 1 provides the optimal policy.

Acknowledgment

The authors wish to thank the anonymous referees for their helpful comments. This work was generously supported by the Japan Society for the Promotion of Science under grant no. JSPS/S-12131.

References

Bäuerle, N. and Rieder, U. (2011) *Markov Decision Processes with Applications to Finance*. Springer-Verlag, Berlin.

Bertsekas, D.P. and Shreve, S.E. (1978) *Stochastic Optimal Control*. Academic Press, New York.

Dufour, F. and Piunovskiy, A. (2010) Multiobjective stopping problem for discrete-time Markov processes: the convex analytic approach. *Journal of Applied Probability* 47: 947–966.

Dynkin, E.B. and Yushkevich A.A. (1979) *Controlled Markov Processes and their Applications*. Springer-Verlag, New York - Berlin.

Easley, D. and Kiefer, N.M. (1988) Controlling a stochastic process with unknown parameters. *Econometrica* 56: 1045–1064.

Ekström, E. and Lu, B. (2011) Optimal selling of an asset under incomplete information. *Int. J. Stoch. Anal.* Art. ID 543590, 17 ,2090–3340.

Ferguson, T.S. (1967) *Mathematical Statistics.* Academic Press, New York - London.

González-Trejo, J.I., Hernández-Lerma, O. and Hoyos-Reyes, L.F. (2003) Minimax control of discrete–time stochastic systems. *SIAM J. Control Optim* 41: 1626–1659.

DeGroot, M.H. (1970) *Optimal Statistical Decisions*. McGraw-Hill Book Co., New York.

van Hee, K.M. (1978) *Bayesian Control of Markov Chains*. Mathematical Centre Tracts, No. 95. Mathematisch Centrum, Amsterdam.

Hernández-Lerma, O. and Marcus, S.I. (1985) Adaptive control of discounted Markov decision chains. *Journal of Optimization Theory and Applications* 46: 227–235.

Hernández-Lerma, O. (1989) *Adaptive Markov Control Processes*, volume 79 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

Hernández-Lerma, O. and Lasserre, J.B. (1996) *Discrete-time Markov Control Processes*. Springer, New York.

Hordijk, A. (1974) *Dynamic Programming and Markov Potential Theory*. Mathematical Centre Tracts, No. 51. Mathematisch Centrum, Amsterdam.

Horiguchi, M. (2001a) Markov decision processes with a stopping time constraint. *Mathematical Methods of Operations Research* 53: 279–295.

Horiguchi, M. (2001b) Stopped Markov decision processes with multiple constraints. *Mathematical Methods of Operations Research* 54: 455–469.

Kurano, M. (1972) Discrete-time Markovian decision processes with an unknown parameter. Average return criterion. *Journal of the Operations Research Society of Japan* 15: 67–76.

Kurano, M. (1983) Adaptive policies in Markov decision processes with uncertain transition matrices. *Journal of Information & Optimization Sciences* 4: 21–40.

Mandl, P. (1974) Estimation and control in Markov chains. *Advances in Applied Probability* 6: 40–60.

Martin, J.J. (1967) *Bayesian Decision Problems and Markov Chains.* Publications in Operations Research, No. 13. John Wiley & Sons Inc., New York.

Piunovskiy, A. B. (2006) Dynamic programming in constrained Markov decision processes. *Control and Cybernetics* **35**: 645–660.

Puterman, M. (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, New York.

Raiffa, H and Schlaifer, R. (1961) *Applied Statistical Decision Theory.* Studies in Managerial Economics. Division of Research, Graduate School of Business Administration, Harvard University, Boston, Mass.

Rieder, U. (1975) Bayesian Dynamic Programming. *Advances in Applied Probability* 7: 330–348.

Ross, S.M. (1970) *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco.

Ross, S.M. (1983) *Introduction to Stochastic Dynamic Programming.* Academic Press, San Diego, CA.

Stadje, W. (1997) An optimal stopping problem with two levels of incomlete information. *Mathem. Methods of Oper. Research* 45: 119–131.

Wald, A. (1950) *Statistical Decision Functions.* John Wiley & Sons Inc., New York.

Wang, X. and Yi, Y. (2009) An optimal investment and consumption model with stochastic returns. *Applied Stoch. Models in Business and Industry* 25: 45–55.

White, D.J. (1969) *Dynamic Programming.* Mathematical Economic Texts, 1. Oliver & Boyd, Edinburgh–London.