MATHEMATICAL CONTROL THEORY
BANACH CENTER PUBLICATIONS, VOLUME 14
PWN – POLISH SCIENTIFIC PUBLISHERS
WARSAW 1985

ON SELF-OPTIMIZING CONTROL OF MARKOV PROCESSES

PETR MANDL

Department of Probability and Mathematical Statistics, Charles University, Prague, Czechoslovakia

1. Introduction

The paper surveys the investigations of the following procedure to control the Markov processes with unknown parameters: The parameters are continuously estimated from the observed trajectory, the optimal stationary policy corresponding to the estimates is found, and the control parameter is adjusted to the value prescribed by that policy ([19] and M. Kurano in [17]). The aim is to approach the optimal stationary policy corresponding to the true parameters. If this is the case, the procedure is called self-optimizing or adaptive. Our survey covers a period of ten years with stress on recent developments and on prospective areas of research.

To explain the basic ideas we consider in § 2 the problem of preventive renewals. § 3 is devoted to finite state Markov processes, § 4 to Markov chains. The last two sections deal with diffusion processes and with special models. The pattern of research is in all cases analogous to the analysis presented here in § 2. Modifications are due to the varying mathematical properties of the models. A consistent denotation of the relevant quantities will help to compare the material in different sections.

2. Renewal processes

1. In the classical model of the renewals we imagine a machine and a particular component of it which is subject to failure. There is an infinite stock of spare components whose life-times (operation times) σ are mutually independent, identically distributed,

$$P(\sigma \leqslant t) = F(t), \quad t \geqslant 0.$$

After failure the component in the machine is instantaneously replaced by a new one. The failure times, coinciding with the replacement times,

form a random sequence of points $T = \{\tau_n,\ n = 1, 2, \ldots\}$ in the time interval $(0, \infty)$. The distribution of the first failure time τ_1 depends on the age of the machine component at time 0. It can therefore be different from F. T is equivalently defined by its *counting process*

$$N_t = \sum_{n=1}^{\infty} \chi\{\tau_n \leqslant t\}, \quad t \geqslant 0.$$

χ is the indicator of the random event in braces. In words, N_t is the number of replacements made until time t.

Assume that F has probability density f, and define the *failure rate* of a component of age $t \geqslant 0$ to be

$$q(t) = f(t)/\bar{F}(t), \quad \text{where} \quad \bar{F}(t) = 1 - F(t) = \int_t^{\infty} f(s)\,ds.$$

We then have

$$P\{\sigma \in (t, t+\Delta) | \sigma > t\} = q(t)\Delta + o(\Delta) \quad \text{as} \quad \Delta \to 0+,$$

provided that q is right-continuous. Return to the renewal process T Define X_t to be the age of the component in operation at time t,

$$X_t = X_0 + t, \quad 0 \leqslant t \leqslant \tau_1, \quad X_t = t - \tau_n, \quad \tau_n < t \leqslant \tau_{n+1}, \quad n = 1, 2, \ldots$$

Let $X^+ = \{X_t^+, t \geqslant 0\}$ be the right-continuous version of X. Since the failure rate depends solely on the age of the component, its evolution in time is

$$Q_t = q(X_t^+), \quad t \geqslant 0. \tag{1}$$

Random function $Q = \{Q_t,\ t \geqslant 0\}$ has the property that the probability of failure in time interval $(t, t+\Delta)$ is $Q_t\Delta + o(\Delta)$ as $\Delta \to 0+$. We shall call Q a failure rate as well. The above-mentioned probability is conditioned on the collection of the events defined on N_s, $s \leqslant t$, in symbols

$$\mathcal{F}_t = \sigma a(N_s,\ 0 \leqslant s \leqslant t).$$

The relation

$$E\{N_{t+\Delta} - N_t | \mathcal{F}_t\} = Q_t\Delta + o(\Delta), \quad \Delta \to 0+,$$

provides the clue to an intuitive proof that

$$M_t = N_t - \int_0^t Q_s\,ds, \quad t \geqslant 0, \tag{2}$$

is a martingale with respect to the increasing system of σ-algebras $\mathcal{F} = \{\mathcal{F}_t,\ t \geqslant 0\}$. In fact, (2) is the *Doob–Meyer decomposition* of submartingale N.

2. Consider now two types of replacements: *service replacements* after failure at cost c_1 and *preventive replacements* before failure at cost c_2, where $c_1 > c_2 > 0$. Both replacements are instantaneous without causing delays. The objective is to find a rule for preventive replacements minimizing the average costs per unit operation time of the machine.

The simpliest rules are the *age replacement policies*. Such policies are given by specifying the replacement age y. When the operation time of the component reaches y, it is replaced by a new one. Let $\Theta(y)$ denote the corresponding average cost per unit time. Obviously, $\Theta(y)$ equals the average cost per replacement divided by the average time between the replacements. Hence,

$$\Theta(y) = \big(c_1 F(y) + c_2 \bar{F}(y)\big)\Big(\int_0^y \bar{F}(s)\,ds\Big)^{-1}.$$

Denote by d the optimal replacement age, and make the following assumptions:

(a) There exists a $d \in (0, \infty)$ such that $\Theta(d) \leqslant \Theta(y)$, $y \in (0, \infty)$.

(b) The components have failure rate $q(t)$ continuous on $[0, \infty)$, and $\lim_{t\to\infty} q(t) = \infty$.

(a) excludes $d = \infty$. (b) implies $\bar{F}(t) = \exp\big\{-\int_0^t q(s)\,ds\big\}$. We shall write briefly Θ for $\Theta(d)$.

The value of d can usually be obtained by equating to zero the derivative of $\Theta(y)$. This is a satisfactory solution provided that the distribution function F is known. If this is not the case, self-optimizing replacement policies are to be used (see J. A. Bather [1]). We shall utilize the relative simplicity of the model to outline a general method for investigating the adaptive controls.

First we shall extend the concept of the age replacement. Under a *general replacement policy* the replacement age is a random function $Z = \{Z_t,\ t \geqslant 0\}$ taking on positive values including $+\infty$. Whenever $X_t = Z_t$, a preventive replacement is made. $T = \{\tau_n,\ n = 1, 2, \ldots\}$ denotes the point process of the replacement times. Introduce the counting process of the service replacements

$${}^1N_t = \sum_{n=1}^{\infty} \chi\{\tau_n \leqslant t,\ X_{\tau_n} < Z_{\tau_n}\}, \qquad t \geqslant 0,$$

and of the preventive replacements

$${}^2N_t = \sum_{n=1}^{\infty} \chi\{\tau_n \leqslant t,\ X_{\tau_n} = Z_{\tau_n}\}, \qquad t \geqslant 0.$$

The counting processes generate a nondecreasing family of σ-algebras

$$\mathcal{F}_t = \sigma a({}^1N_s, {}^2N_s, s \leqslant t), \quad t \geqslant 0.$$

Z_t depends on the history up to time t. Thus, Z_t should be $\mathcal{F}_t$-measurable, $t \geqslant 0$. Moreover, we assume that the trajectories of Z are left-continuous and piecewise continuous.

Let the basic space Ω be the set of all possible sample paths of $({}^1N, {}^2N)$ equipped with σ-algebra $\mathcal{F}_\infty$. The probability distribution of the replacement process under policy $Z = \{Z_t, t \geqslant 0\}$ is a probability measure P^Z on $(\Omega, \mathcal{F}_\infty)$ with the following properties:

(a) P^Z-almost surely we have: $X_t = Z_t$ if and only if ${}^2N_t - {}^2N_{t-} = 1$.

(b) ${}^1M_t = {}^1N_t - \int_0^t q(X_s)ds$, $t \geqslant 0$, is a martingale with respect to $\mathcal{F} = \{\mathcal{F}_t, t \geqslant 0\}$ on $(\Omega, \mathcal{F}_\infty, P^Z)$.

(a) comes from the definition of the preventive replacement age. (b) was explained in Subsection 1.

3. Since c_1 is the cost of an after-failure replacement and c_2 the cost of a preventive replacement, the total cost incurred until time t is

$$C_t = c_1 {}^1N_t + c_2 {}^2N_t.$$

Next we are going to prove that Θ, the minimal average cost under the age replacement policies, cannot be improved by using general policies, i.e., that for arbitrary Z

$$\varliminf_{t\to\infty} t^{-1}C_t \geqslant \Theta \text{ a.s.;} \tag{3}$$

a.s. stands for almost surely. It is a common fact in Markovian decision problems that the optimum is achieved on stationary policies. However, the method of proof will enable us to proceed further.

To prove (3) we try to find a bounded function $w(x)$, $x \in [0, \infty)$ such that

$$S_t = C_t - t\Theta + w(X_t^+), \quad t \geqslant 0, \tag{4}$$

is under each policy Z a submartingale with respect to $\mathcal{F}$. If we succeed, we have the decomposition

$$S_t = M_t + A_t, \quad t \geqslant 0, \tag{5}$$

where M is a martingale, A a nonnegative nondecreasing process. From the law of large numbers for M we then get

$$0 = \lim_{t\to\infty} t^{-1}M_t \leqslant \varliminf_{t\to\infty} t^{-1}S_t = \varliminf_{t\to\infty} t^{-1}C_t - \Theta \text{ a.s.,}$$

which is the desired result. Note that in (4)

$$X_t^+ = X_t \chi\{N_t = N_{t-}\}.$$

Hence, S_t is $\mathscr{F}_t$-measurable. S has right-continuous trajectories.

To obtain $w(x)$, or to prove its existence, let us take the viewpoint of the impulsive control theory. A policy Z in fact specifies a sequence of stopping times $\{\sigma_n,\ n = 1, 2, \ldots\}$ at which the trajectory $\{X_t,\ t \geqslant 0\}$ is shifted into 0 for cost c_2. Consider, as is done in the optimal impulsive control theory, the expected discounted cost, i.e.,

$$E_x^Z \int_0^\infty e^{-lt} dC_t.$$

$l > 0$ is the discount factor, x denotes the age of the machine component when the process starts, $X_0 = x$. The "quasi-variational inequalities" for the infimum

$$u(x) = \inf_Z E_x^Z \int_0^\infty e^{-lt} dC_t$$

are

$$u'(x) + q(x)\big(c_1 + u(0) - u(x)\big) - lu(x) \geqslant 0, \tag{6}$$

$$c_2 + u(0) - u(x) \geqslant 0, \tag{7}$$

$$\big(u'(x) + q(x)\big(c_1 + u(0) - u(x)\big) - lu(x)\big)\big(c_2 + u(0) - u(x)\big) = 0. \tag{8}$$

Let us sketch the demonstration of (6)–(8). It consists in replacing the stopping by the killing with a rate not exceeding a level $\bar{r}$, and in letting $\bar{r} \to \infty$. Thus, consider the replacement process assuming that the controller selects a rate $R = \{R_t,\ t \geqslant 0\}$, $0 \leqslant R_t \leqslant \bar{r}$, so that

$$P_x^R({}^2N_{t+\Delta} - {}^2N_t > 0 \mid \mathscr{F}_t) = R_t \Delta + o(\Delta), \qquad \Delta \to 0+.$$

In other words, P_x^R is the probability measure on $(\Omega, \mathscr{F}_\infty)$ for which

$${}^1N_t - \int_0^t q(X_s)\,ds, \qquad {}^2N_t - \int_0^t R_s\,ds, \qquad t \geqslant 0,$$

are martingales with respect to $\mathscr{F}$. Denote

$$\bar{u}(x) = \inf_R E_x^R \int_0^\infty e^{-lt} dC_t.$$

The Bellman equation for $\bar{u}(x)$ is derived by the usual argument. We have

$$\bar{u}(x) = \inf_{0 \leqslant r \leqslant \bar{r}} \big[(1 - q(x)\Delta - r\Delta)e^{-l\Delta}\bar{u}(x + \Delta) + q(x)\Delta\big(c_1 + \bar{u}(0)\big) + r\Delta\big(c_2 + \bar{u}(0)\big) + o(\Delta)\big], \qquad \Delta \to 0+.$$

Hence,

$$(9)\quad \bar{u}'(x)+q(x)\left(c_1+\bar{u}(0)-\bar{u}(x)\right)+\inf_{0<r\leqslant\bar{r}} r\left(c_2+\bar{u}(0)-\bar{u}(x)\right)-l\,\bar{u}(x)=0.$$

Further, since stopping is killing with infinite rate,

$$(10)\quad \lim_{\bar{r}\to\infty}\bar{u}(x)=u(x).$$

(6) follows from (9), (10), because the third term in (9) is always non-positive. To verify (7) note that its contrary

$$c_2+u(0)-u(x)<0$$

implies that the third term in (9) tends to $-\infty$. Finally, to get (8) we observe that if

$$c_2+u(0)-u(x)>0,$$

then the third term in (9) equals zero for $\bar{r}$ sufficiently large.

The required function $w(x)$ is obtained, on the heuristic level of reasoning, by a standard passage from the discounted cost to the average cost, letting $l\to 0+$. We have

$$\lim_{l\to 0+} lu(x)=\Theta,$$

and set

$$\lim_{l\to 0+}\left(u(x)-u(0)\right)=w(x).$$

From (6)–(8) follow the inequalities

$$(11)\quad w'(x)+q(x)\left(c_1-w(x)\right)-\Theta\geqslant 0,$$

$$(12)\quad c_2-w(x)\geqslant 0,$$

$$(13)\quad \left(w'(x)+q(x)\left(c_1-w(x)\right)-\Theta\right)\left(c_2-w(x)\right)=0.$$

Moreover, $w(0)=0$.

The subsequent proposition yields decomposition (5).

PROPOSITION 1. *Let $w(x)$, $x\in[0,\infty)$, be bounded, continuously differentiable, and such that* (11)–(13) *hold. Then*

$$(14)\quad C_t-t\Theta+w(X_t^+)=M_t+A_t,\qquad t\geqslant 0,$$

where

$$(15)\quad M_t=w(x_0)+\int_0^t\left(c_1-w(X_s)\right)\left(d\,{}^1N_s-q(X_s)\,ds\right),\qquad t\geqslant 0,$$

$$A_t = \int_0^t \big(w'(X_s) + q(X_s)\,(c_1 - w(X_s)) - \Theta\big)\,ds + \int_0^t (c_2 - w(X_s))\,d\,{}^2N_s, \quad t \geqslant 0. \tag{16}$$

M is a martingale, A is a nonnegative nondecreasing process.

The proof follows from the relation

$$\int_0^t w'(X_s)\,ds = w(X_t^+) - w(X_0) + \int_0^t w(X_s)\,d({}^1N_s + {}^2N_s).$$

It has been pointed out by J. Zabczyk that (11)–(13) give an idea how to treat the optimal impulsive control problems under the average cost criterion, a question listed under the unsolved ones in [27] by M. Robin.

4. Equality holds in (11) for $x \in [0, d]$ and $w(d) = c_2$, where d is the optimal replacement age. If $Z_t = d$, $t \geqslant 0$, then $A_t = A_0$, $t \geqslant 0$. Thus, (14) and

$$\lim_{t\to\infty} t^{-1} M_t = 0 \text{ a.s.}$$

imply

$$\lim_{t\to\infty} t^{-1} C_t = \Theta \text{ a.s.} \tag{17}$$

Under a *self-optimizing* (sequentially improving) policy d is reached in the limit, i.e.,

$$\lim_{t\to\infty} Z_t = d \text{ a.s.} \tag{18}$$

Whenever (18) holds, (17) is true, and M fulfils the law of the iterated logarithm

$$\overline{\lim_{t\to\infty}} \pm M_t / \sqrt{(2t \log\log t)} = \sigma \text{ a.s.}, \tag{19}$$

where

$$\sigma^2 = \int_0^d (c_1 - w(y))^2\,dF(y) \Big/ \int_0^d \bar{F}(y)\,dy.$$

There are too many procedures satisfying (18). Therefore the results concerning the speed of convergence in (17) serve to classify the procedures. From (14) and (19) the following results are obtainable. (See [25] with a different decomposition (14) giving A nondecreasing only for q nondecreasing.)

PROPOSITION 2. *If*

$$\overline{\lim_{t\to\infty}} |Z_t - d| \sqrt{(t/\log\log t)} < \infty \text{ a.s.}, \tag{20}$$

then

$$\overline{\lim_{t\to\infty}} \pm (C_t - t\Theta)/\sqrt{(2t\log\log t)} = \sigma \text{ a.s.} \tag{21}$$

PROPOSITION 3. *If* $0 \leqslant r < 1/2$, *and*

$$\overline{\lim_{t\to\infty}} |Z_t - d| t^r < \infty \text{ a.s.}, \tag{22}$$

then

$$\lim_{t\to\infty} (C_t - t\Theta) t^{r-1} = 0 \text{ a.s.}$$

The proof consists in demonstrating the negligibility of A under (20) and in estimating its growth under (22). (21) is an upper bound for the convergence speed of $t^{-1}C_t$ to Θ. It is reached if $Z_t = d$, $t \geqslant 0$. Proposition 2 states that self-optimizing policies satisfying (20) are in the sense of (21) indistinguishable from the optimal policy. It will be seen in the next subsection that (20) can be attained under certain regularity conditions if f is specified up to unknown parameters. The same is true for Bather's procedure ([1]) with an appropriate choice of the truncation probabilities. On the basis of the Kiefer–Wolfowitz stochastic approximation method a replacement procedure has been constructed ([25]) for which (22) holds with any $r < 1/4$.

5. We come to the subject proper of this survey, namely to the procedure based on inserting a parameter estimate into the optimal replacement age (see V. Menyhértová [26]). Suppose that the failure distribution, and hence the rate q, is specified up to a parameter a, the true value a_0 of which is unknown to us. The rate

$$q(t, a), \quad t \geqslant 0, \quad a \in \mathcal{A},$$

is given. To each a there corresponds an optimal replacement age $d(a)$. We let

$$Z_t = d(a_t^*), \quad t \geqslant 0, \tag{23}$$

where a_t^* is the maximum likelihood estimate of a_0 on the basis of $({}^1N_s, {}^2N_s)$, $s \leqslant t$. The log-likelihood function to be maximized is

$$L_t(a) = \int_0^t \log q(X_s, a)\, d\,{}^1N_s - \int_0^t q(X_s, a)\, ds, \quad a \in \mathcal{A}.$$

The investigation starts with proving, under suitable hypotheses, the strong consistency of $\{a_t^*,\ t \geqslant 0\}$, i.e.,

$$\lim_{t\to\infty} a_t^* = a_0 \text{ a.s.} \tag{24}$$

for an arbitrary policy Z. The existence of a function $l(a)$, $a \in \mathcal{A}$, is established such that $l(a_0) = 0$, $l(a) < 0$, $a \neq a_0$, and

$$\overline{\lim_{t\to\infty}}\, t^{-1}\big(L_t(a) - L_t(a_0)\big) \leqslant l(a) \text{ a.s.} \tag{25}$$

for arbitrary Z. (25) is analogous to (3), and the method presented in Subsection 3 works in the proof. (25) implies that a_t^* cannot be far from a_0 for large t, and hence (24) follows.

Having (24) and assuming $d(a)$ continuous in a one readily gets for the policy given by (23)

$$\lim_{t\to\infty} Z_t = d(a_0) \text{ a.s.}, \tag{26}$$

or the property of being self-optimizing.

Further asymptotic properties of $\{a_t^*,\ t \geqslant 0\}$ under (23), (26) are obtained from Taylor's development of $L_t'(a)$ at a_0. Its one-dimensional version is

$$L_t'(a_0) = (a_t^* - a_0)L_t''(a_0) + \int_0^{a_t^*} \big(L_t''(a) - L_t''(a_0)\big)\,da. \tag{27}$$

Having verified the law of the iterated logarithm for the martingale $\{L_t'(a_0),\ t \geqslant 0\}$, the strong law of large numbers for $\{L_t''(a_0),\ t \geqslant 0\}$, and the negligibility of the last term in (27), one gets

$$\lim_{t\to\infty} |a_t^* - a_0| \sqrt{(t/\log\log t)} = \text{const} < \infty \text{ a.s.} \tag{28}$$

Provided that $d(a)$ is Lipschitz continuous, (28) yields (20) with $d = d(a_0)$.

By applying the central limit theorem to $\{L_t'(a_0), t \geqslant 0\}$, conditions are derived for the asymptotic normality of $\{a_t^*,\ t \geqslant 0\}$.

3. Finite state Markov processes

The probability distribution of a Markov process $X = \{X_t,\ t \geqslant 0\}$ with finite state space I is specified by the initial distribution $P(X_0 = i)$, $i \in I$, and by the *transition rates* $q(i, j, t)$, which for $i \neq j$ have the following meaning:

$$P(X_{t+\Delta} = j \mid X_t = i) = q(i, j, t)\Delta + o(\Delta), \qquad \Delta \to 0+.$$

We set

$$-\sum_{j \neq i} q(i, j, t) = q(i, i, t), \quad i \in I.$$

Let the basic space Ω be the set of paths (trajectories) $\omega(t)$, $t \geqslant 0$, with values in I, piecewise constant and right-continuous. Set $X_t(\omega) = \omega(t)$. X generates a nondecreasing family of σ-algebras

$$\mathscr{F}_t = \sigma a(X_s,\ s \leqslant t), \quad t \geqslant 0.$$

Take any state $i \in I$, and consider the process $\chi\{X_t = i\}$, $t \geqslant 0$. A martingale is obtained by subtracting from $\chi\{X_t = i\}$ the integral of the transition rate.

PROPOSITION 4. *X on $(\Omega, \mathscr{F}_\infty, P)$ is a Markov process with piecewise continuous transition rates $q(i, j, t)$, $i, j \in I$, $t \geqslant 0$, if and only if*

$${}^{i}M_t = \chi\{X_t = i\} - \int_0^t q(X_s, i, s)\, ds, \quad t \geqslant 0,\ i \in I,$$

are martingales with respect to $\mathscr{F}$.

The proof of Proposition 4 involves the forward system of Kolmogorov differential equations. The martingale characterization of the probability distribution is suitable for controlled Markov processes.

The dynamics of a *controlled Markov process* with state space I is defined by the transition rates

$$q(i, j; z), \quad i, j \in I,$$

depending on a control parameter $z \in J$. Let J be a compact set, and let q be continuous in z. We limit ourselves to the time-homogeneous case. The initial distribution $P(X_0 = i) = p_i$, $i \in I$, is assumed to be fixed.

The control parameter at time t is chosen on the basis of the observation until time t, i.e., on the basis of the events from $\mathscr{F}_t$. A *control* is a random function $Z = \{Z_t,\ t \geqslant 0\}$ on $(\Omega, \mathscr{F}_\infty)$ with values in J, left-continuous and such that Z_t is $\mathscr{F}_t$-measurable for $t \geqslant 0$. Closed loop controls

$$Z_t = \bar{z}(X_t^-), \quad t \geqslant 0, \tag{29}$$

are called *stationary* or *homogeneous Markovian controls*. In (29) $\bar{z}(i)$ is a mapping from I into J, $\{X_t^-,\ t \geqslant 0\}$ is the left-continuous version of X.

With any control Z we want to associate a probability measure P^Z on $(\Omega, \mathscr{F}_\infty)$, being the probability distribution of X under the control Z. If (29) holds, we define P^Z so that X is a Markov process with transition rates $q(i, j; \bar{z}(i))$, $i, j \in I$. From Proposition 4 it follows that then

$${}^{i}M_t = \chi\{X_t = i\} - \int_0^t q(X_s, i; Z_s)\, ds, \quad t \geqslant 0,\ i \in I, \tag{30}$$

are martingales with respect to $\mathscr{F}$. The extension to general controls is evident.

DEFINITION 1. The *probability distribution* of the controlled process X under the control Z is a probability measure P^Z on $(\Omega, \mathscr{F}_\infty)$ such that ${}^iM$, $i \in I$, are martingales with respect to $\mathscr{F}$, and $P^Z(X_0 = i) = p_i$, $i \in I$.

P^Z exists and is unique.

Next we introduce an evaluation of the trajectories called here the cost. By the *cost until time* t we understand the integral

$$C_t = \int_0^t c(X_s, Z_s)ds, \quad t \geqslant 0. \tag{31}$$

$c(i, z)$ is a given continuous function on $I \times J$. The aim is to minimize the average cost per unit time. We consider first stationary controls (29), and make the following hypothesis:

ASSUMPTION 1. For each $\bar{z}$, under control (29) the states of X are recurrent and mutually communicating.

Assumption 1 implies that under (29) there exist limit probabilities

$$\lim_{t\to\infty} P^Z(X_t = i) = p_i^\infty(\bar{z}) > 0, \quad i \in I.$$

Denote

$$\Theta(\bar{z}) = \sum_i p_i^\infty(\bar{z})c(i, \bar{z}(i)).$$

From the law of large numbers for Markov processes follows

$$\lim_{t\to\infty} t^{-1}C_t = \Theta(\bar{z}) \ P^Z\text{-a.s.}$$

The minimal average cost is

$$\inf_{\bar{z}} \Theta(\bar{z}) = \Theta(\hat{z}) = \Theta.$$

$\Theta(\bar{z})$, being a continuous function on a compact set, takes on its minimal value.

To be able to proceed as in § 2 we seek a function $w(j)$, $j \in I$, such that

$$S_t = C_t - t\Theta + w(X_t), \quad t \geqslant 0,$$

is a submartingale (with respect to $\mathscr{F}$) for each Z. By Definition 1 for arbitrary w,

$$M_t = \sum_j w(j)\,{}^jM_t = w(X_t) - \int_0^t \sum_j q(X_s, j; Z_s)w(j)ds, \ t \geqslant 0,$$

is a martingale for each Z. Setting

$$f(i,z) = c(i,z) + \sum_j q(i,j;z)w(j) - \Theta, \quad i \in I,\ z \in J, \tag{32}$$

we have

$$C_t - t\Theta + w(X_t) = M_t + \int_0^t f(X_s, Z_s)ds, \quad t \geqslant 0. \tag{33}$$

If we achieve

$$f(i,z) \geqslant 0, \quad i \in I,\ z \in J, \tag{34}$$

then (33) is the Doob–Meyer decomposition of submartingale S. But (34) holds if

$$f(i, \hat{z}(i)) = 0, \quad i \in I,$$

as follows from an application of the law of large numbers to (33).

(33) enables us to obtain results analogous to those presented in Subsection 4 of § 2 in the problem formulation as well as in the fact that martingale limit theorems are applied to M in the proofs (see [21]).

The occurrence of an unknown parameter a in the transition rates leads to the problem of combining the parameter estimation and the process control treated in Subsection 5 of § 2. Thus, given the transition rates

$$q(i,j;z,a), \quad i,j \in I,\ z \in J,\ a \in \mathscr{A},$$

we associate with each a the mapping $\hat{z}(i,a)$, $i \in I$, producing an optimal stationary control, and set

$$Z_t = \hat{z}(X_t, a_t^*), \quad t \geqslant 0. \tag{35}$$

$\{a_t^*,\ t \geqslant 0\}$ is an estimate of the true value a_0. Controls (35) based on the estimate maximizing the log-likelihood function

$$L_t(a) = \sum_{s \leqslant t} \chi\{X_s^- \neq X_s\} q(X_s^-, X_s; Z_s, a) + \int_0^t q(X_s, X_s; Z_s, a)ds, \quad a \in \mathscr{A},$$

or, more generally, on a minimum contrast estimate have been investigated. Another class of estimates considered are discrete time recursive estimates. The central limit theorem and the law of the iterated logarithm hold for C_t as $t \to \infty$ under (35) with the same parameters as under the optimal stationary control

$$Z_t = \hat{z}(X_t, a_0), \quad t \geqslant 0,$$

provided that certain regularity conditions are fulfilled (see [21]). The cost functional (31) can be generalized to include also the cost arising from the jumps of the trajectory.

4. Markov chains

To arrive at the controlled Markov chains $\{X_n, n = 0, 1, \ldots\}$ we only have to replace in § 3 the continuous time by a discrete parameter $n = 0, 1, \ldots$, and the transition rates by the transition probabilities. The counterparts to (32) and (33) are (see [22])

$$f(i, z) = c(i, z) + \sum_j p(i, j; z) w(j) - w(i) - \Theta, \quad i \in I, \ z \in J,$$

$$C_N - N\Theta + w(X_N) = M_N + \sum_{n=0}^{N-1} f(X_n, Z_n), \quad N = 0, 1, \ldots \tag{36}$$

Most of the papers on our subject deal with the Markov chains. Let us review the directions of the work done.

If the state space I is countably infinite, additional hypotheses are to be made to prevent the trajectories from running too fast to infinity. A. Hordijk ([10]) introduced a Liapunov type condition, which, together with certain continuity properties of the transition matrices, ensures the existence of an average cost optimal control. It is as follows: For some $k \in I$ and a nonnegative sequence $\{y(j), j \in I\}$ we have

$$|c(i, z)| + 1 + \sum_{j \neq k} p(i, j; z) y(j) \leqslant y(i), \quad i \in I, \ z \in J.$$

Conditions of this kind were used in [24] to ensure the validity of the limit laws for the martingale M, to obtain estimates of the remaining terms in (36), and hence to make it possible to proceed as in the finite case. Several contributions to this method were made in [11] by M. Kolonko, who dealt with the control and the parameter estimation in countable semi-Markov processes. See also M. Kolonko and M. Schäl [12] for a more general approach.

J. P. Georgin in [9] treats Markov chains with a general state space.

Further papers concentrate on the estimation of the unknown parameter with the aim either to weaken the conditions for the consistency of the maximum likelihood and the minimum contrast estimates (V. Borkar, P. Varaiya in [3], [4], B. T. Doshi, S. E. Shreve in [7], P. R. Kumar [13], P. R. Kumar, A. Becker [14], P. R. Kumar, W. Lin [15], or to study a special case in detail (B. Sagalovsky in [28]). J. Brychta's thesis [5] contains properties of the recursive estimates.

5. Diffusion processes

Paper [20] is on diffusion processes, but a systematic study of the one-dimensional case was made by V. Lánská-Dufková in [8]. Let $X = \{X_t, t \geqslant 0\}$ fulfil the Itô stochastic differential equation

$$dX_t = a(X_t, Z_t)dt + b(X_t)dW_t, \quad t \geqslant 0. \tag{37}$$

$W = \{W_t, t \geqslant 0\}$ is a Wiener process, $a(x, z), b(x), x \in (-\infty, \infty), z \in J$, are given functions. As in § 3, the cost until time t is defined by (31).

Recall that the differential generator of a finite state Markov process is the matrix of its transition rates, while the differential generator of (37) is

$$D = \frac{1}{2} b(x)^2 \frac{d^2}{dx^2} + a(x, z)\frac{d}{dx}.$$

Applying this observation to (32), one gets

$$f(x, z) = c(x, z) + Dw(x) - \Theta, \quad x \in (-\infty, \infty), \ z \in J.$$

Function $w(x)$ is to be such that (34) holds. From the Itô formula for stochastic differentials follows (33) with M being a local martingale. Since M has continuous quadratic variation, it can be transformed to a Wiener process by means of a random change of the time scale. Thus, the limit properties of M can be deduced from those of the Wiener process.

V. Lánská deals also with the case where (37) has the form

$$dX_t = \big(a_1(X_t, Z_t) + \alpha a_2(X_t)\big)dt + b(X_t)dW_t, \quad t \geqslant 0,$$

where α is an unknown parameter estimated by the maximum likelihood method and by a recursive method derived from continuous stochastic approximations. Minimum contrast estimates were introduced by her in [18].

6. Special models

B. T. Doshi investigates in [6] an inventory system with a controlled deterministic input. The incoming demands form a compound Poisson process. Their sizes are exponentially distributed. The two parameters of the demand process are unknown.

The adaptive control of the discrete time linear systems with a quadratic cost is approached in [23] by the methods presented in this review.

P. Kunderová in [16] deals with finite state Markov processes in which instantaneous shifts of the trajectory can be made.

References

[1] J. A. Bather, *On the sequential construction of an optimal age replacement policy*, in: Proc. 41st Session ISI, New Delhi 1977.

[2] A. Bensoussan, *Optimal impulsive control theory*, Lecture Notes Control and Inform. Sci. 16, Springer-Verlag, 1979, 17–41.

[3] V. Borkar, P. Varaiya, *Adaptive control of Markov chains. I. Finite parameter set*, IEEE Trans. Automatic Control 24 (1979), 953–957.

[4] —, —, *Adaptive control of Markov chains*, in: Lecture Notes Control and Inform. Sci. 16, Springer-Verlag, 1979, 294–296.

[5] J. Brychta, *Recursive parameter estimates in controlled Markov chains*, Thesis, Charles University, Prague 1977 [in Czech].

[6] B. T. Doshi, *Adaptive control of a production-inventory system*, J. Appl. Probability 18 (1981), 204–215.

[7] B. T. Doshi, S. E. Shreve, *Strong consistency of a modified maximum likelihood estimator for controlled Markov chains*, J. Appl. Probability 17 (1980), 726–734.

[8] V. Dufková, *On controlled one-dimensional diffusion processes with unknown parameter*, Advances in Appl. Probability 9 (1977), 105–124.

[9] J. P. Georgin, *Estimation et contrôle des chaînes de Markov sur des espaces arbitraires*, in: Lecture Notes Math. 636, Springer-Verlag, 1978, 71–113.

[10] A. Hordijk, *Dynamic programming and Markov potential theory*, Math. Centrum, Amsterdam 1974.

[11] M. Kolonko, *Dynamische Optimierung unter Unsicherheit in einem Semi-Markoff-Modell mit abzählbarem Zustandsraum*, Thesis, Friedrich-Wilhelm University, Bonn 1980.

[12] M. Kolonko, M. Schäl, *Optimal control of semi-Markov chains under uncertainty with applications to queueing models*, Proc. Operations Res. 9 (1980), 430–435.

[13] P. R. Kumar, *Adaptive control with a compact parameter set*, Math. Res. Report 80–16, Univ. of Maryland, Baltimore County 1980.

[14] P. R. Kumar, A. Becker, *A new family of optimal adaptive controllers*, Math. Res. Report 80–18, Univ. of Maryland, Baltimore County, 1980.

[15] P. R. Kumar, W. Lin, *Optimal adaptive controllers for unknown systems*, Math. Res. Report 80–21, Univ. of Maryland, Baltimore County 1980.

[16] P. Kunderová, *Markov processes with renewals*, Thesis, Charles University, Prague 1977 [in Czech].

[17] M. Kurano, *Discrete-time Markovian decision processes with an unknown parameter—average return criterion*, J. Operations Res. Soc. Japan 15 (1972), 67–76.

[18] V. Lánská, *Minimum contrast estimation in diffusion processes*, J. Appl. Probability 16 (1979), 65–75.

[19] P. Mandl, *On the control of a Markov chain in the presence of unknown parameters*, in: Trans. 6th Prague Conf. Inform. Theory, etc. 1971, Academia, Prague 1973, 601–612.

[20] —, *An application of Itô's formula to stochastic control systems*, Lecture Notes Math. 294, Springer-Verlag, 1972, 8–13.

[21] —, *Asymptotically optimal controls of Markov processes involving unknown parameters*, in: Proc. Prague Symp. Asymptotic Stat. 1973, Vol. II, Charles University, Prague 1974, 247–256.

[22] —, *Estimation and control in Markov chains*, Advances in Appl. Probability 6 (1974), 40–60.

[23] P. Mandl, *The use of optimal stationary policies in the adaptive control of linear systems*, in: Proc. Symp. J. Neyman 1974, PWN, Warsaw 1977, 223–242.

[24] —, *On the adaptive control of countable Markov chains*, in: Probability Theory, Banach Center Publ. 5, PWN, Warsaw 1979, 159–173.

[25] —, *On the sequential improvement of replacement policies*, in: Proc. 2nd Prague Symp. Asymptotic Stat. 1978, JČSMF Prague and North-Holland, 1979, 277–290.

[26] V. Menyhértová, *On adaptive replacement policies*, Kybernetika (Prague) 16 (1980), 512–525.

[27] M. Robin, *Contrôle impulsionnel des processus de Markov*, Thesis, University Paris IX, Paris 1978.

[28] B. Sagalovsky, *Adaptive control and parameter estimation in Markov chains: A linear case*, IEEE Trans. Automatic Control.