Acta Universitatis Carolinae. Mathematica et Physica

Petr Mandl; Gerhard Hübner

Transient phenomena and self-optimizing control of Markov chains

Acta Universitatis Carolinae. Mathematica et Physica, Vol. 26 (1985), No. 1, 35--51

Persistent URL: `http://dml.cz/dmlcz/142545`

Transient Phenomena and Self-optimizing Control of Markov Chains

PETR MANDL

Department of Probability and Mathematical Statistics, Charles University*)

GERHARD HÜBNER

Institute for Mathematical Stochastics, Hamburg University**)

Received 10 October 1984

Finite state controlled Markov chains with transition probabilities depending on a parameter are considered. The parameter values converge to a limit. The best asymptotic distribution of the reward is found, and conditions are given under which the optimum obtains for adaptive controls.

Pojednává se o konečných řízených Markovových řetězcích s pravděpodobnostmi přechodu závislými na parametru. Hodnoty parametru konvergují k limitě. Je nalezeno nejlepší asymptotické rozložení výnosu a udány podmínky, za nichž se dosáhne optima při adaptivních řízeních.

Исследуются конечные управляемые цепи Маркова с вероятностями перехода зависящими от параметра. Значения параметра стремятся к пределу. Найдено наилучшее асымптотическое поведение дохода и прэдложены условия, при которых оптимум достигается для адаптивных управлений.

0. Summary

In [2] and in subsequent papers (see [3], [4] for review) self-optimizing controls of Markov processes were studied, which are based on the principle of inserting the estimates of the unknown parameters into the optimal stationary control. This is in fact a sequential application of the certainty equivalence principle. In the present paper we investigate the situation, when the parameters of the system change with time, and converge to a limit. Thus, we assume the occurence of a transient phenomenon, and investigate its influence on the performance of the system. In the cases where the transient behaviour is known, partially known or unknown conditions are given for the optimality of suitable (self-optimizing) controls and the asymptotic normality of the pertinent rewards.

To make the presentation as simple as possible, we consider, as in [2], the case of a finite Markov chain. Moreover, we concentrate on parameter estimates by the

*) Sokolovská 83, 186 00 Praha 8, Czechoslovakia

**) Bundesstrasse 55, 2 Hamburg 13, Federal Republic of Germany maximum likelihood method although the results can be extended in a straightforward way to other minimum contrast methods. Finally, to shorten the Taylor expansions, we suppose that the unknown parameter is one-dimensional.

The paper is aimed to present some novel aspects of nonstationary decision processes treated in [1].

1. Fundamentals

Consider a system S taking on states from a finite set say I. S is observed at times $n = 0, 1, 2, \dots$. Let X_n denote the state of S at time n. The transition law of S is specified by *transition probabilities* from state i to state k

$$(1) \qquad p(i, k; z, \alpha), \quad z \in \boldsymbol{Z}, \quad \alpha \in A, \quad i, k \in I.$$

z is a control parameter ranging in a closed bounded set $\boldsymbol{Z} \subset R^q$. α is a parameter specifying the conditions of the transition. We shall denote by β_n its value for the passage from X_n to X_{n+1}, $n = 0, 1, \dots$. Let $A \subset R^1$ be closed and bounded.

The value of the control parameter at time n is a random variable depending on the observed trajectory

$$(2) \qquad Z_n = z_n(X_0, \dots, X_n), \quad n = 0, 1, \dots.$$

The control $Z = \{Z_n, n = 0, 1, \dots\}$ is called stationary if

$$(3) \qquad Z_n = \mathbf{z}(X_n), \quad n = 0, 1, \dots,$$

where $\mathbf{z}$ is a mapping from I to $\boldsymbol{Z}$. We write then $Z \sim \mathbf{z}$. Let the distribution of the initial state be fixed,

$$P(X_0 = i) = p_i, \quad i \in I.$$

From the above said it follows that under (2) the probability distribution P^Z of $X = \{X_n, n = 0, 1, \dots\}$ satisfies

$$(4) \qquad P^Z(X_{n+1} = k \mid X_0 = i_0, \dots, X_n = i_n) = p(i_n, k; z_n(i_0, \dots, i_n), \beta_n).$$

To evaluate the performance of the system we introduce the quantity

$$C_N = \sum_{n=0}^{N-1} c(X_n, X_{n+1}; Z_n, \beta_n), \quad N = 1, 2, \dots.$$

C_N will be called *the reward* up to time N. The transition probabilities (1) as well as the functions defining the reward from one transition

$$c(i, k; z, \alpha), \quad z \in \boldsymbol{Z}, \quad \alpha \in A, \quad i, k \in I,$$

are assumed to be continuous in (z, α). We make also the following hypothesis.

Assumption 1. For each $\alpha \in A$ and each stationary control $\mathbf{z}$ the matrix $\|p(i, k; \mathbf{z}(i), \alpha)\|_{i,k \in I}$ is indecomposable.

Next we restate some facts about stationary controls. Assume (3) and

$$\beta_n = \alpha\,, \quad n = 0, 1, 2, \ldots .$$

Then X is a time homogeneous Markov chain which is recurrent in virtue of Assumption 1. Denote by $\pi_i(\mathbf{z}, \alpha)$, $i \in I$, its stationary distribution. The expectation

$$\Theta(\mathbf{z}, \alpha) = \sum_i \pi_i(\mathbf{z}, \alpha) \sum_k p(i, k; \mathbf{z}(i), \alpha)\, c(i, k; \mathbf{z}(i), \alpha)$$

is the stationary reward per one step under control $\mathbf{z}$. It can be shown that $\Theta(\mathbf{z}, \alpha)$ is continuous in $\mathbf{z}$. The maximal stationary reward

$$\Theta(\alpha) = \sup_{\mathbf{z}} \Theta(\mathbf{z}, \alpha)$$

is thus attained for a control which we shall denote $\mathbf{z}(i, \alpha)$, $i \in I$. The following characterization of $\Theta(\alpha)$ due to R. Bellman is important.

Proposition 1. For $\alpha \in A$, $\Theta(\alpha)$ is the unique number such that auxiliary constants

(5) $$w_i(\alpha)\,, \quad i \in I\,,$$

can be found that

(6) $$\sup_{z \in \mathbf{Z}} \left[\sum_k p(i, k; z, \alpha)\,(c(i, k; z, \alpha) + w_k(\alpha)) - w_i(\alpha) - \Theta(\alpha)\right] = 0\,, \quad i \in I\,.$$

The constants (5) are determined uniquely up to a shift by a constant. $\mathbf{z}(i, \alpha)$ is optimal if and only if the supremum in (6) is attained for $z = \mathbf{z}(i, \alpha)$, $i \in I$.

The expression in square brackets in (6) will be denoted by

$$\Phi(i, z, \alpha)\,, \quad i \in I\,, \quad z \in \mathbf{Z}\,, \quad \alpha \in A\,.$$

Thus we have

(7) $$\sup_{z \in \mathbf{Z}} \Phi(i, z, \alpha) = 0 = \Phi(i, \mathbf{z}(i, \alpha), \alpha)\,, \quad i \in I\,.$$

Further we state an assumption pertaining to the consistency of the maximum likelihood method.

Assumption 2. For $i, k \in I$ either $p(i, k; z, \alpha) > 0$ for $z \in \mathbf{Z}$, $\alpha \in A$, or $p(i, k; z, \alpha) = 0$, $z \in \mathbf{Z}$, $\alpha \in A$. If $\alpha, \alpha' \in A$, $\alpha \neq \alpha'$, then for each $\mathbf{z}$

$$\|p(i, k; \mathbf{z}(i), \alpha)\|_{i,k \in I} \neq \|p(i, k; \mathbf{z}(i), \alpha')\|_{i,k \in I}\,,$$

i.e., the transition laws are different.

Finally we recall the most simple versions of the law of large numbers and the central limit theorem for martingales.

Proposition 2. Let

$$M_N = \sum_{n=0}^{N-1} Y_n\,, \quad N = 1, 2, \ldots,$$

be a martingale with respect to a nondecreasing sequence of Borel fields $\{\mathscr{F}_N, N = 0, 1, \ldots\}$. Let

$$|Y_n| \leqq \text{const. a.s.}, \quad n = 0, 1, \ldots .$$

Then

$$\lim_{N\to\infty} N^{-1} M_N = 0 \quad \text{a.s.} \tag{8}$$

If, in addition,

$$\lim_{N\to\infty} N^{-1} \sum_{n=0}^{N-1} E\{Y_n^2 \mid \mathscr{F}_n\} = \zeta \quad \text{in prob.}, \tag{9}$$

where ζ is a constant, then $M_N/\sqrt{N}$ has asymptotically normal distribution $N(0, \zeta)$ as $N \to \infty$.

a.s. is abbreviation for almost surely. The proofs of Propositions 1, 2 are given in [2].

2. General results for transitory behaviour

In the rest of the paper we assume that the parameter sequence β_n, $n = 0, 1, \ldots,$ is induced by a transient phenomenon, i.e.,

$$\lim_{n\to\infty} \beta_n = \alpha_0 . \tag{10}$$

C_N as $N \to \infty$ is to be made as large as possible. The main instrument in our investigation will be the martingale introduced in the next proposition.

Proposition 3. Under arbitrary control Z the sequence

$$\begin{aligned} M_N = C_N - \sum_{n=0}^{N-1} \Theta(\beta_n) + w_{X_n}(\beta_{N-1}) - w_{X_0}(\beta_0) + \\ + \sum_{n=1}^{N-1} (w_{X_n}(\beta_{n-1}) - w_{X_n}(\beta_n)) - \sum_{n=0}^{N-1} \Phi(X_n, Z_n, \beta_n), \quad N = 1, 2, \ldots, \end{aligned} \tag{11}$$

is a martingale with respect to the Borel fields $\mathscr{F}_N$ of random events defined in terms of $X_0, X_1, \ldots, X_N$, $N = 0, 1, \ldots$.

Proof. We have

$$M_N = \sum_{n=0}^{N-1} Y_n , \quad N = 1, 2, \ldots,$$

with

$$\begin{aligned} Y_n = c(X_n, X_{n+1}; Z_n, \beta_n) - \Theta(\beta_n) + w_{X_{n+1}}(\beta_n) - w_{X_n}(\beta_n) - \\ - \Phi(X_n, Z_n, \beta_n) = c(X_n, X_{n+1}; Z_n, \beta_n) + w_{X_{n+1}}(\beta_n) - \end{aligned} \tag{12}$$

$$- \sum_k p(X_n, k; Z_n, \beta_n) \left(c(X_n, k; Z_n, \beta_n) + w_k(\beta_n)\right).$$

M_N is $\mathscr{F}_N$-measurable, and according to (4)

$$E^Z\{Y_n \mid \mathscr{F}_n\} = 0, \quad n = 0, 1, \ldots \quad \square$$

Note that in virtue of the hypotheses made the functions

$$\Theta(\alpha), \quad w_i(\alpha) - w_j(\alpha), \quad \Phi(i, z, \alpha), \quad i, j \in I,$$

are continuous in α and in (z, α), respectively. Hence, they are bounded. Next proposition states that $\Theta(\alpha_0)$ is the asymptotic upper bound for the average reward attainable.

Proposition 4. Under arbitrary control Z

$$\overline{\lim_{N \to \infty}} N^{-1} C_N \leqq \Theta(\alpha_0) = \lim_{N \to \infty} N^{-1} \sum_{n=0}^{N-1} \Theta(\beta_n), \quad P^Z \quad \text{a.s.} \tag{13}$$

Proof. Note that by (7) the last sum in (11) is always nonpositive. The before last sum is $o(N)$ as $N \to \infty$, since

$$w_{X_n}(\beta_{n-1}) - w_{X_n}(\beta_n) \to 0$$

because of (10). Thus, dividing (11) by N we get (13) from (8) as $N \to \infty$. $\square$

The limiting evolution of S is characterized by the parameter value α_0. The fixed optimal stationary control corresponding to α_0 was denoted $\mathbf{z}(\cdot, \alpha_0)$. It seems and proves to be best for the controller to approach this control in the limit, i.e., to have

$$\lim_{n \to \infty} (Z_n - \mathbf{z}(X_n, \alpha_0)) = 0. \tag{14}$$

The self-optimizing controls dealt with in Section 3 serve to this aim. Here we state basic relations holding for controls satisfying (14).

Proposition 5. Let Z be such that (14) holds almost surely. Then

$$\lim_{N \to \infty} N^{-1} C_N = \Theta(\alpha_0) \tag{15}$$

P^Z almost surely.

Proof. Recalling the proof of Proposition 4 and with regard to (7) we see that (15) follows from

$$0 = \lim_{n \to \infty} (\Phi(X_n, Z_n, \beta_n) - \Phi(X_n, \mathbf{z}(X_n, \alpha_0), \alpha_0)) = \lim_{n \to \infty} \Phi(X_n, Z_n, \beta_n) \; P^Z \text{ a.s.} \quad \square$$

Remark 1. If (14) holds in probability then we have (15) with convergence in probability.

Before stating the next proposition let us compute $E^Z\{Y_n^2 \mid \mathscr{F}_n\}$ for the martingale difference (12). We have from (4)

$$E^Z\{Y_n^2 \mid \mathscr{F}_n\} = c_2(X_n; Z_n, \beta_n), \quad n = 0, 1, \dots,$$

where

$$c_2(i, z, \alpha) = \sum_k p(i, k; z, \alpha)\,(c(i, k; z, \alpha) + w_k(\alpha))^2 -$$

$$-\,(\sum_k p(i, k; z, \alpha)\,(c(i, k; z, \alpha) + w_k(\alpha)))^2, \quad i \in I, \quad z \in \boldsymbol{Z}, \quad \alpha \in A.$$

We recall also that $\pi_i(\mathbf{z}(\cdot, \alpha_0), \alpha_0)$, $i \in I$, is the limiting distribution for the Markov chain X under the stationary control $\mathbf{z}(\cdot, \alpha_0)$.

Proposition 6. Let $w_i(\alpha)$, $i \in I$, be continuously differentiable, and let

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} |\beta_{n+1} - \beta_n| = 0. \tag{16}$$

Then for any Z satisfying

$$P^Z\text{-}\lim_{n\to\infty} (Z_n - \mathbf{z}(X_n, \alpha_0)) = 0 \tag{17}$$

we have

$$\lim_{N\to\infty} P^Z\left(\frac{C_N - \sum_{n=0}^{N-1} \Theta(\beta_n)}{\sqrt{N}} < x\right) \geqq \Psi\left(\frac{x}{\sqrt{\zeta}}\right), \quad x \in (-\infty, \infty), \tag{18}$$

where $\Psi(x/\sqrt{\zeta})$ is the distribution function of the law $N(0, \zeta)$ with

$$\zeta = \sum_i \pi_i(\mathbf{z}(\cdot, \alpha_0), \alpha_0)\, c_2(i; \mathbf{z}(i, \alpha_0), \alpha_0).$$

P^Z-lim means convergence in probability.

In terms of stochastic ordering of random variables (18) states that C_N is asymptotically stochastically smaller than $N(\sum_{n=0}^{N-1} \Theta(\beta_n), N\zeta)$. Hence, Proposition 6 contains for large N a characterization of the best possible performance of the system when random fluctuations of the actual value of C_N are taken into consideration.

Proof of Proposition 6. Let the hypotheses and (17) hold. Consider (11). We have

$$P^Z\text{-}\lim_{N\to\infty} N^{-1} \sum_{n=0}^{N-1} E^Z\{Y_n^2 \mid \mathscr{F}_n\} = P^Z\text{-}\lim_{N\to\infty} N^{-1} \sum_{n=0}^{N-1} c_2(X_n, Z_n, \beta_n) = \zeta. \tag{19}$$

(19) is verified by the same reasoning as that which led to Remark 1. Hence, by Proposition 2, $M_N/\sqrt{N}$ is asymptotically $N(0, \zeta)$ as $N \to \infty$.

Further,

$$w_{X_n}(\beta_{n-1}) - w_{X_n}(\beta_n) = O(|\beta_{n-1} - \beta_n|) \quad \text{as} \quad n \to \infty \,.$$

Consequently, in virtue of (16),

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{n=1}^{N-1} (w_{X_n}(\beta_{n-1}) - w_{X_n}(\beta_n)) = 0 \,. \tag{20}$$

The last sum in (11) is nonpositive. Thus we have for $x \in (-\infty, \infty)$

$$P^Z\left(\frac{C_N - \sum_{n=0}^{N-1} \Theta(\beta_n)}{\sqrt{N}} < x\right) \geqq P^Z((M_N - w_{X_N}(\beta_{N-1}) + \tag{21}$$

$$+ w_{X_0}(\beta_0) - \sum_{n=1}^{\mathrm{N}-1} (w_{X_n}(\beta_{n-1}) - w_{X_n}(\beta_n)))/\sqrt{(N)} < x) \,.$$

The right-hand side tends to $\Psi(x/\sqrt{\zeta})$ as $N \to \infty$. This establishes (18). □

3. Control with full information

Corollary 1. Under the hypotheses of Proposition 6 set

$$Z_n = \mathbf{z}(X_n, \beta_n) \,, \quad n = 0, 1, 2, \dots \,. \tag{22}$$

Let $\mathbf{z}(i, \alpha)$, $i \in I$, be continuous at α_0. Then (17) holds because of (10). Since also

$$\sum_{n=1}^{N-1} \Phi(X_n, Z_n, \beta_n) = 0 \,, \quad N = 1, 2, \dots \,,$$

we have equality in (21). Consequently,

$$\lim_{N\to\infty} P^Z\left(\frac{C_N - \sum_{n=0}^{N-1} \Theta(\beta_n)}{\sqrt{N}} < x\right) = \Psi\left(\frac{x}{\sqrt{\zeta}}\right), \quad x \in (-\infty, \infty) \,.$$

Remark 2. Note that (16) is fulfilled if

$$\beta_n \geqq \beta_{n+1} \,, \quad n = 0, 1, \dots \,.$$

Consider next the case that $\{\beta_n, n = 0, 1, \dots\}$ satisfies

$$\beta_{2m} \geqq \alpha_0 \geqq \beta_{2m+1} \,, \quad m = 0, 1, \dots \,, \quad |\beta_n - \alpha_0| \geqq |\beta_{n+1} - \alpha_0| \,, \quad n = 0, 1, \dots \,.$$

We can speak about damped oscillations of the parameter. Two step transitions are to be considered if (16) does not hold. As example take a Markov chain X with two states and two control parameter values, say 0,1. Let the transition probability matrix be

$$\begin{pmatrix} \frac{1}{2} + (-1)^z \alpha, & \frac{1}{2} - (-1)^z \alpha \\ \frac{1}{2} + (-1)^z \alpha, & \frac{1}{2} - (-1)^z \alpha \end{pmatrix}, \quad z = 0, 1, \quad |\alpha| \leqq \tfrac{1}{3}.$$

Further let

$$c(0, k; z, \alpha) = \alpha, \quad c(1, k; z, \alpha) = 0, \quad k = 0, 1, \quad z = 0, 1, \quad |\alpha| \leqq \tfrac{1}{3}.$$

Optimal control for α fixed is to take always $z = 0$, and this yields the average reward

$$\Theta(\alpha) = (\tfrac{1}{2} + \alpha)\,\alpha.$$

Assume

$$\alpha_0 = 0, \quad \beta_{2m} \geqq \beta_{2m+2}, \quad \beta_{2m+1} = -\beta_{2m}, \quad m = 0, 1, \ldots.$$

Introduce a Markov chain $\bar{X}$ one step of which corresponds to two transitions in the original chain for α equal β and $-\beta$, respectively. It can be seen that the control identically equal to 1 is optimal, and leads to

$$\bar{\Theta}(\beta) = 2\beta^2.$$

Applying Proposition 6 to $\bar{X}$, and noting that $\zeta = 0$, we obtain

$$\lim_{N \to \infty} P^Z\left(\frac{C_N - \sum_{n=0}^{N-1} \beta_n^2}{\sqrt{N}} < x\right) = 1, \quad x > 0. \tag{23}$$

(23) holds for arbitrary Z since for $\alpha_0 = 0$ all controls are optimal.

For

$$Z_n = 1, \quad n = 0, 1, \ldots,$$

we have also

$$\lim_{N \to \infty} P^Z\left(\frac{C_N - \sum_{n=0}^{N-1} \beta_n^2}{\sqrt{N}} < x\right) = 0, \quad x < 0. \tag{24}$$

On the other hand, (22) means

$$Z_n = 0, \quad n = 0, 1, \ldots \tag{25}$$

(24) does not hold under (25) for $\beta_{2m} = \frac{1}{3}(m + 1)^{-1/6}$ e.g.

Next we consider the situation when the value α_0 is unknown. We shall distinguish two cases: 1. The controller has some information about the transitory sequence $\{\beta_n, n = 0, 1, \ldots\}$. 2. The controller is not aware of the transient phenomenon or neglects it.

4. Control with partial information

Let

$$\beta_n = \alpha_0 + b_0\, g(n), \quad n = 0, 1, \ldots, \tag{26}$$

where

$$\lim_{n\to\infty} g(n) = 0 .$$

The sequence $\{g(n),\ n = 0, 1, \ldots\}$ is known to the controller, while the constants α_0, b_0 are unknown to him. $\alpha_0 \in A$, $b_0 \in B$, A and B closed and bounded. To maximize the average reward the controller would like to employ (22). Since he does not know α_0 b_0, he sets

$$Z_n = \mathbf{z}(X_n, \alpha_n^* + b_n^* g(n)), \quad n = 0, 1, \ldots, \tag{27}$$

where α_n^*, b_n^* are maximum likelihood estimates of α_0, b_0. I.e.,

$$L_n(\alpha_n^*, b_n^*) = \sup_{\alpha\in A, b\in B} L_n(\alpha, b), \qquad n = 1, 2, \ldots, \tag{28}$$

$$L_n(\alpha, b) = \sum_{m=0}^{n-1} \ln p(X_m, X_{m+1};\ Z_m, \alpha + b\, g(m)) .$$

Proposition 7. Let Z be an arbitrary control. Let (α_n^*, b_n^*), $n = 1, 2, \ldots$, be pairs of random variables satisfying (28). Then for

$$\beta_n^* = \alpha_n^* + b_n^* g(n), \quad n = 0, 1, \ldots,$$

holds

$$\lim_{n\to\infty} \beta_n^* = \alpha_0 \quad P^Z \text{ a.s.} \tag{29}$$

With regard to Proposition 3 the proof of (29) is the same as the proof given in [2] for $g(n) = 0$, $n = 0, 1, \ldots$.

Corollary 2. Let $\mathbf{z}(i, \alpha)$, $i \in I$, be continuous at α_0, and let (27) hold. Then

$$\lim_{n\to\infty} (Z_n - \mathbf{z}(X_n, \alpha_0)) = 0 \quad P^Z \text{ a.s.} \tag{30}$$

Therefore Propositions 5 and 6 apply.

(30) enables us to use Taylor's expansion to derive additional properties of the estimates. For the rest of the paper we make the following hypothesis.

Assumption 3. α_0 is an interior point of A. The second derivatives with respect to α,

$$p''(i, k; z, \alpha), \quad i, k \in I, \quad z \in \mathbf{Z},$$

exist in a neighbourhood of α_0, and are continuous in (z, α) at $(\mathbf{z}(i, \alpha_0), \alpha_0)$. Moreover,

$$J = \sum_i \sum_k \pi_i(\mathbf{z}(\cdot, \alpha_0), \alpha_0) \frac{p'(i, k; \mathbf{z}(i, \alpha_0), \alpha_0)^2}{p(i, k; \mathbf{z}(i, \alpha_0), \alpha_0)} > 0 .$$

We say that

$$(\beta_N^* - \beta_N)\sqrt{N}\,, \quad N = 0, 1, \ldots, \tag{31}$$

is *almost bounded in mean square*, if to arbitrary $\varepsilon > 0$ there exist a random event D and a constant K so that

$$P^Z(D) > 1 - \varepsilon\,, \tag{32}$$

$$NE^Z\chi_D(\beta_N^* - \beta_N)^2 \leqq K\,, \quad N = 0, 1, \ldots \tag{33}$$

χ_D is the indicator of D.

Lemma 1. Let (27) hold and let $\mathbf{z}(i, \alpha)$, $i \in I$, be continuous at α_0. Further let

$$g(n) = \frac{h(n)}{n^g}\,, \quad n = 1, 2, \ldots, \quad 0 < g < \tfrac{1}{2}\,, \tag{34}$$

where

$$\lim_{n \to \infty} n\left(\frac{h(n+1)}{h(n)} - 1\right) = 0\,. \tag{35}$$

Then (31) is almost bounded in mean square.

Proof. Note that (35) implies

$$\lim_{n \to \infty} n^\varepsilon h(n) = \infty\,, \quad \lim_{n \to \infty} n^{-\varepsilon} h(n) = 0 \quad \text{for} \quad \varepsilon > 0\,.$$

For

$$R_N = \sum_{n=1}^{N} \frac{1}{n^g} \sim \frac{1}{1 - g} N^{1-g}\,, \quad N \to \infty\,, \tag{36}$$

we have

$$\sum_{n=1}^{N} g(n) = \sum_{n=1}^{N-1} R_n(h(n) - h(n+1)) + R_N\, h(N)\,.$$

(35) and (36) imply that the sum on the right is $o(\sum_{n=1}^{N} g(n))$. Hence,

$$\sum_{n=1}^{N} g(n) \sim \frac{1}{1 - g} N\, g(N)\,, \quad N \to \infty\,. \tag{37}$$

By a similar argument

$$\sum_{n=1}^{N} n(g(n) - g(n+1)) = \sum_{n=1}^{N} h(n)\left(\frac{n}{n^g} - \frac{n}{(n+1)^g}\right) + \tag{38}$$

$$+ \sum_{n=1}^{N} \frac{n}{(n+1)^g}(h(n) - h(n+1)) \sim g \sum_{n=1}^{N} g(n) \sim \frac{g}{1 - g} N\, g(N)\,, \quad N \to \infty\,.$$

To simplify the Taylor expansion of $L_N(\alpha_N^*, b_N^*)$ assume $p''(i, k; z, \alpha)$, $i, k \in I$, continuous on $\boldsymbol{Z} \times A$. and write shortly

$$\ln p' = \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln p(X_n, X_{n+1}; Z_n, \alpha)\big|_{\alpha=\beta_n} \tag{39}$$

$$\ln p'' = \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} \ln p(X_n, X_{n+1}; Z_n, \alpha)\big|_{\alpha=\tilde{\beta}_n} \tag{40}$$

with $\tilde{\beta}_n$ between $\alpha_N^* + b_N^* g(n)$ and β_n. Note that $\ln p'$ and $\ln p''$ depend on n which is suppressed for convenience. Then

$$\frac{1}{N}(L_N(\alpha_N^*, b_N^*) - L_N(\alpha_0, b_0)) = \frac{1}{N}\sum_{n=0}^{N-1} (\alpha_N^* + b_N^* g(n) - \alpha_0 - b_0\, g(n)) \ln p' + \tag{41}$$

$$+ \frac{1}{2N}\sum_{n=0}^{N-1} (\alpha_N^* + b_N^* g(n) - \alpha_0 - b_0\, g(n))^2 \ln p'' \,.$$

The last term is a quadratic form in

$$\alpha_N^* - \alpha_0\,, \quad (b_N^* - b_0)\, g(N)$$

the matrix of which is

$$\begin{pmatrix} \dfrac{1}{2N}\displaystyle\sum_{n=0}^{N-1} \ln p''\,, & \dfrac{1}{2N\, g(N)}\displaystyle\sum_{n=0}^{N-1} g(n) \ln p''\,. \\ \dfrac{1}{2N\, g(N)}\displaystyle\sum_{n=0}^{N-1} g(n) \ln p''\,, & \dfrac{1}{2N\, g(N)^2}\displaystyle\sum_{n=0}^{N-1} g(n)^2 \ln p'' \end{pmatrix}. \tag{42}$$

Let us investigate the limit of (42) as $N \to \infty$. Set

$$S_N = \sum_{n=0}^{N-1} \ln p'' \,.$$

With regard to (29) and by a similar argument as in the proof of Proposition 5 we get

$$\lim_{N\to\infty} \frac{1}{N} S_N = \sum_i \sum_k \pi_i(\mathbf{z}(\cdot\ \alpha_0), \alpha_0)\, p(i, k; \mathbf{z}(i, \alpha_0), \alpha_0)\,. \tag{43}$$

$$.\, (\ln p)''\,(i, k; \mathbf{z}(i, \alpha_0), \alpha_0) = -J \quad P^{\mathrm{Z}} \text{ a.s.}$$

Further

$$\sum_{n=0}^{N-1} g(n) \ln p'' = \sum_{n=1}^{N-1} (g(n-1) - g(n))\, S_n + g(N-1)\, S_N \,.$$

Hence from (38), (43),

$$\lim_{N\to\infty} \frac{1}{N\, g(N)} \sum_{n=0}^{N-1} g(n) \ln p'' = -\left(\frac{g}{1-g} + 1\right) J = \frac{-J}{1-g} \quad P^{\mathrm{Z}} \text{ a.s.}$$

Similarly,

$$\lim_{N \to \infty} \frac{1}{N\, g(N)^2} \sum_{n=0}^{N-1} g(n)^2 \ln p'' = \frac{-J}{1 - 2g} \quad P^Z \text{ a.s.}$$

We conclude that (42) converges almost surely as $N \to \infty$ to the negative definite matrix

$$-\frac{J}{2} \begin{pmatrix} 1 & , & \dfrac{1}{1-g} \\ \dfrac{1}{1-g} & , & \dfrac{1}{1-2g} \end{pmatrix}.$$

This implies that for N sufficiently large the last term in (41) does not exceed $-K_0 \Delta_N^2$, where K_0 is a positive constant, and

$$\Delta_N^2 = (\alpha_N^* - \alpha_0)^2 + (b_N^* - b_0)^2\, g(N)^2\,.$$

Since the left-hand side of (41) is nonnegative, we conclude that there exists an almost surely finite random variable v such that

$$(44) \qquad 0 \leqq (\alpha_N^* - \alpha_0) \frac{1}{N} \sum_{n=0}^{N-1} \ln p' + (b_N^* - b_0)\, g(N)\,.$$

$$\cdot \frac{1}{N\, g(N)} \sum_{n=0}^{N-1} g(n) \ln p' - K_0 \Delta_N^2\,, \quad N = v\,, \quad v + 1, \ldots.$$

To establish that (31) is almost bounded in mean square take $\varepsilon > 0$ arbitrary. Find N_0 such that

$$P^Z(v \leqq N_0) > 1 - \varepsilon\,,$$

and set $D = \{v \leqq N_0\}$. From (44) follows applying Schwarz's inequality

$$(45) \qquad 0 \leqq \frac{1}{N} \sqrt{[E^Z \chi_D (\alpha_N^* - \alpha_0)^2]}\, \sqrt{[\sum_{n=0}^{N-1} E^Z (\ln p')^2]} +$$

$$+ \frac{1}{N} \sqrt{[E^Z \chi_D (b_N^* - b_0)^2\, g(N)^2]} \cdot \sqrt{\left[\frac{1}{g(N)^2} \sum_{n=0}^{N-1} g(n)^2\, E^Z (\ln p')^2\right]} - K_0 E^Z \chi_D \Delta_N^2$$

$$N = N_0, N_0 + 1, \ldots.$$

We have used the fact that $\ln p'$ are martingale differences. From (45) we obtain

$$0 \leqq K_1 \sqrt{[N E^Z \chi_D \Delta_N^2]} - K_0 N E^Z \chi_D \Delta_N^2\,, \quad N = N_0, N_0 + 1, \ldots$$

for a suitable constant K_1. But this implies that

$$N E^Z \chi_D \Delta_N^2\,, \quad N = 1, 2, \ldots$$

is bounded. Since

$$2NE^{Z}\chi_D\Delta_N^2 \geqq NE^{Z}\chi_D(\beta_N^* - \beta_N)^2 ,$$

we have (33). □

The subsequent proposition asserts that the control (27) is as good as (22) in the sense introduced in Section 2.

Proposition 8. Under the hypotheses of Lemma 1 let

(46) $$w_i(\alpha) , \quad i \in I , \quad \Phi(i, \mathbf{z}(i, \alpha), \beta) , \quad i \in I , \quad \beta \in A ,$$

have continuous derivative with respect to $\alpha \in A$. Then for the control

$$Z_n = \mathbf{z}(X_n, \alpha_n^* + b_n^* g(n)) , \quad n = 0, 1, \dots ,$$

holds

(47) $$\lim_{N\to\infty} P^{Z}\left(\frac{C_N - \sum_{n=0}^{N-1} \Theta(\beta_n)}{\sqrt{N}} < x\right) = \Psi\left(\frac{x}{\sqrt{\zeta}}\right), \quad x \in (-\infty, \infty) .$$

Proof. Consider the proof of Proposition 6. (17) holds by Corollary 2. We shall show that (18) can be strengthened to (47). Note that

$$\sum_{n=0}^{\infty} |\beta_{n+1} - \beta_n| = |b_0| \sum_{n=0}^{\infty} |g(n + 1) - g(n)| < \infty .$$

Thus (16), and hence (20), is fulfilled. To demonstrate (47) it suffices to establish

(48) $$P^{Z}\text{-}\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \Phi(X_n, Z_n, \beta_n) = 0 .$$

For β in the interior of A we conclude from

$$\Phi(i, \mathbf{z}(i, \beta), \beta) = 0 , \quad \frac{\mathrm{d}}{\mathrm{d}\alpha} \Phi(i, \mathbf{z}(i, \alpha), \beta)|_{\alpha=\beta} = 0 , \quad i \in I ,$$

that

(49) $$-\Phi(i, \mathbf{z}(i, \alpha), \beta) \leqq |\alpha - \beta| \, \varphi(\alpha - \beta) , \quad i \in I , \quad \alpha \in A ,$$

with

$$\lim_{x\to 0} \varphi(x) = 0 .$$

φ can be assumed bounded.

(48) is valid if

(50) $$\lim_{N\to\infty} E^{Z}\left(\chi_D \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \Phi(X_n, Z_n, \beta_n)\right) = 0$$

with D having probability arbitrarily close to 1. Let D be such that (32), (33) hold.

Then

$$-E^Z\left(\chi_D \frac{1}{\sqrt{N}}\sum_{n=1}^{N-1}\Phi(X_n, Z_n, \beta_n)\right) \leqq \frac{1}{\sqrt{N}}\sum_{n=1}^{N-1} E^Z(\chi_D|\beta_n^* - \beta_n|\,\varphi(\beta_n^* - \beta_n)) \leqq$$

$$\leqq \frac{1}{\sqrt{N}}\sum_{n=1}^{N-1}\sqrt{[E^Z\chi_D(\beta_n^* - \beta_n)^2]}\,\sqrt{[E^Z\varphi(\beta_n^* - \beta_n)^2]} \leqq \tag{51}$$

$$\leqq \frac{1}{\sqrt{N}}\sum_{n=1}^{N-1}\frac{K}{\sqrt{n}}\sqrt{[E^Z\varphi(\beta_n^* - \beta_n)^2]}\,.$$

Since

$$\lim_{n\to\infty}(\beta_n^* - \beta_n) = 0 \quad P^Z \text{ a.s.}\,,$$

we have

$$\lim_{n\to\infty} E^Z\varphi(\beta_n^* - \beta_n)^2 = 0\,.$$

Hence (51) implies (50). □

5. Control without information on the transitory behaviour

Case 1

We shall show that the controller may neglect the transient phenomenon if

$$\beta_n - \alpha_0 = O\left(\frac{1}{\sqrt{n}}\right), \quad n \to \infty\,, \tag{52}$$

which is the situation complementary to (34) assumed in the first part of this section. The result can be extended under additional smoothness hypotheses.

Let the control be performed as follows. The controller assumes

$$\beta_n = \alpha_0\,, \quad n = 0, 1, \dots,$$

and employs

$$Z_n = \mathbf{z}(X_n, \alpha_n^*)\,, \quad n = 0, 1, \dots, \tag{53}$$

where α_n^* is his maximum likelihood estimate of α_0. Namely,

$$L_n(\alpha_n^*) = \sup_{\alpha\in A} L_n(\alpha)\,, \quad n = 1, 2, \dots, \tag{54}$$

$$L_n(\alpha) = \sum_{m=0}^{n-1}\ln p(X_m, X_{m+1}; Z_m, \alpha)\,.$$

Let us stress that the true situation is given by the sequence $\{\beta_n, n = 0, 1, \dots\}$ satisfying (10). Corollary 2 remains valid with (27) replaced by (53). The following analogue of Proposition 7 is valid

Proposition 9. Let Z be an arbitrary control. Let α_n^*, $n = 1, 2, \ldots$, be a sequence of random variables satisfying (54). Then

$$\lim_{n\to\infty} \alpha_n^* = \alpha_0 \quad P^Z \text{ a.s.}$$

Lemma 2. Let (16), (52), (53) hold, and $\mathbf{z}(i, \alpha)$, $i \in I$, be continuous at α_0. Then

$$(\alpha_N^* - \beta_N)\sqrt{N}\,, \quad N = 0, 1, \ldots\,, \tag{55}$$

is almost bounded in mean square.

Proof. Suppose again for brevity that $p''(i, k; z, \alpha)$, $i, k \in I$, is continuous on $Z \times A$. Retain the denotations (39), (40) with $\tilde{\beta}_n$ between α_0 and β_n, and introduce

$$\ln p_0' = \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln p(X_n, X_{n+1}; Z_n, \alpha)\big|_{\alpha=\alpha_0}\,,$$

$$\ln p_0'' = \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} \ln p(X_n, X_{n+1}; Z_n, \alpha)\big|_{\alpha=\tilde{\alpha}_n}$$

where $\tilde{\alpha}_n$ lies between α_0 and α_n^*. Then

$$\frac{1}{N}(L_N(\alpha_N^*) - L_N(\alpha_0)) = (\alpha_N^* - \alpha_0)\frac{1}{N}\sum_{n=0}^{N-1} \ln p_0' + (\alpha_N^* - \alpha_0)^2 \frac{1}{2N}\sum_{n=0}^{N-1} \ln p_0''\,. \tag{56}$$

With regard to (54) expansion (56) can be used to establish the almost mean square boundedness of

$$(\alpha_N^* - \alpha_0)\sqrt{N}\,, \quad N = 0, 1, \ldots\,, \tag{57}$$

in the same way as (41) in the proof of Lemma 1. However, there is the distinction that $\ln p_0'$ are no longer martingale differences. Thus to estimate $E^Z(\sum \ln p_0')^2$ we proceed as follows.

$$E^Z(\sum_{n=0}^{N-1} \ln p_0')^2 = E^Z(\sum_{n=0}^{N-1} (\ln p' + (\alpha_0 - \beta_n) \ln p''))^2 \leqq \tag{58}$$

$$\leqq 2\sum_{n=0}^{N-1} E^Z(\ln p')^2 + 2E^Z(\sum_{n=0}^{N-1} (\alpha_0 - \beta_n) \ln p'')^2\,.$$

Since

$$|\sum_{n=0}^{N-1} (\alpha_0 - \beta_n) \ln p''| \leqq \text{const.} \sqrt{N}\,, \quad N = 1, 2, \ldots\,, \tag{59}$$

we have

$$E^Z(\sum_{n=0}^{N-1} \ln p_0')^2 \leqq \text{const.}\, N\,, \quad N = 1, 2, \ldots\,,$$

which is the inequality needed to infere as in the proof of Lemma 1 that (57) is almost bounded in mean square. (55) has the same property because of (52). □

The next statement is an exact analogue of Proposition 8.

Proposition 10. Under the hypotheses of Lemma 2 let (46) have continuous derivative with respect to $\alpha \in A$. Then (47) holds for the control (53).

Case 2

Assume now

$$\beta_n = \alpha_0 + b_0\, g(n)\,, \quad n = 1, 2, \ldots, \tag{60}$$

with $g(n)$ as in (34) and (35). Instead of (59) we have by (37) the estimate

$$|\sum_{n=0}^{N-1} (\alpha_0 - \beta_n) \ln p''| \leqq \text{const.}\, N\, g(N)\,, \quad N = N_0, N_0 + 1, \ldots .$$

Thus, from (58)

$$E^Z(\sum_{n=0}^{N-1} \ln p_0')^2 \leqq \text{const.}\, N^2\, g(N)^2\,, \quad N = N_0, N_0 + 1, \ldots .$$

From here it is deduced using the method of the proof of Lemma 2 that

$$0 \leqq \text{const.}\, \sqrt{[E^Z \chi_D (\alpha_N^* - \alpha_0)^2]}\, |g(N)| - \\ - \text{const.}\, E^Z \chi_D (\alpha_N^* - \alpha_0)^2\,, \quad N = N_0, N_0 + 1, \ldots .$$

The last constant is positive, and $P^Z(D)$ can be arbitrarily close to 1. But this implies

$$E\chi_D(\alpha_N^* - \alpha_0)^2 \leqq \text{const.}\, g(N)^2\,, \quad N = N_0, N_0 + 1, \ldots,$$

and hence also

$$E\chi_D(\alpha_N^* - \beta_N)^2 \leqq \text{const.}\, g(N)^2\,, \quad N = N_0, N_0 + 1, \ldots . \tag{61}$$

Proposition 11. Let (60) hold with

$$g(n) = \frac{h(n)}{n^g}\,, \quad n = 1, 2, \ldots, \quad \tfrac{1}{4} < g < \tfrac{1}{2}\,, \tag{62}$$

and $h(n)$ satisfying (35). Let $\mathbf{z}(i, \alpha)$, $i \in I$, be continuous at α_0, let $w_i(\alpha)$, $i \in I$, have a continuous derivative, and

$$\Phi(i, \mathbf{z}(i, \alpha), \beta)\,, \quad i \in I\,, \quad \beta \in A\,,$$

two continuous derivatives with respect to $\alpha \in A$. Then (47) is fulfilled for the control (53).

Proof. (48) is to be verified. Instead of (49) we have the estimate

$$- \Phi(i, \mathbf{z}(i, \alpha), \beta) \leqq \text{const.}\, (\alpha - \beta)^2\,, \quad i \in I\,, \quad \alpha, \beta \in A\,.$$

Hence, from (61),

$$- E^Z\left(\chi_D \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \Phi(X_n, Z_n, \beta_n)\right) \leqq \frac{\text{const.}}{\sqrt{N}} \sum_{n=0}^{N-1} E^Z \chi_D (\alpha_n^* - \beta_n)^2 \leqq \frac{\text{const.}}{\sqrt{N}} \sum_{n=0}^{N-1} g(n)^2\,,$$

$$N = N_0 + 1, N_0 + 2, \dots .$$

The last expression converges to 0 as $N \to \infty$ because of (62). This proves (48). □

Corollary 3. Propositions 10 and 11 are valid with (53) replaced by

$$Z_n = \mathbf{z}(X_n, \alpha_0), \quad n = 0, 1, \dots ,$$

as it is seen by inspecting the respective proofs.

References

[1] FEDERGRUEN, A., SCHWEITZER, P. J.: Nonstationary Markov decision problems with converging parameters. J. Optimization Theory and Appl. 34 (1981), 207—241.

[2] MANDL, P.: Estimation and control in Markov chains. Adv. Appl. Probability 6 (1974), 40—60.

[3] MANDL, P.: On self-optimizing control of Markov processes, Banach Center Publ. Vol. 14, Math. Control Theory, PWN, Warsaw 1983, 347—362.

[4] SCHÄL, M.: Asymptotic results for sequential Markov decision models under uncertainty, Statistics and Decisions 2 (1984), 39—62.