KYBERNETIKA — VOLUME 9 (1973), NUMBER 4

A Connection between Controlled Markov Chains and Martingales

PETR MANDL

A finite controlled Markov chain is considered. It is assumed that its control converges to a stationary one. The law of large numbers and the central limit theorem for the reward are obtained from the theory of martingales.

1. INTRODUCTION

The mathematical model studied in the present paper consists of two sequences of random variables $\{X_n, Z_n, n = 0, 1, \ldots\}$. The state variables X_n, $n = 0, 1, \ldots$, take on values from a finite set $\boldsymbol{I}$. The control variables are functions of the observed state variables

$$Z_n = z_n(X_0, \ldots, X_n), \quad n = 0, 1, \ldots$$

The values of Z_n belong to $\boldsymbol{Z}(X_n)$, $n = 0, 1, \ldots$, where the sets $\boldsymbol{Z}(j)$, $j \in \boldsymbol{I}$, are supposed to be closed and bounded in $\boldsymbol{R}^s$. The functions $\{z_n(j_0, \ldots, j_n), n = 0, 1, \ldots\}$ constitute the controler's policy, briefly the control. The control is called stationary if $z_n(j_0, \ldots, j_n) = z(j_n)$, $n = 0, 1, \ldots$ We make the hypothesis that the double sequence $\{X_n, Z_n, n = 0, 1, \ldots\}$ is a controlled Markov chain, i.e. that

$$(1) \quad P(X_{n+1} = k \mid X_n, Z_n, X_{n-1}, \ldots, X_0, Z_0) = p(X_n, k; Z_n), \quad k \in \boldsymbol{I}, n = 0, 1, \ldots,$$

where $p(j, k; z)$, $z \in \boldsymbol{Z}(j)$, $j, k \in \boldsymbol{I}$, are the transition probabilities from state j into state k, when the control parameter equals z. The performance of the system is measured by the quantity

$$C_M = \sum_{n=0}^{M-1} c(X_n, X_{n+1}; Z_n), \quad M = 1, 2, \ldots,$$

called the reward from the chain up to time M. We assume that the functions

$$p(j, k; z),\ c(j, k; z),\ z \in \mathbf{Z}(j),\ j, k \in \mathbf{I},$$

are continuous on $\mathbf{Z}(j)$, $j \in \mathbf{I}$.

Let Θ be a real number. We shall study the properties of $C_M - M\Theta$ for $M \to \infty$. To do this, we add to $C_M - M\Theta$ a correcting term in order to obtain a martingale. We introduce auxiliary constants w_j, $j \in \mathbf{I}$, and set

$$\varphi(j, z) = \sum_k p(j, k; z)\,[c(j, k; z) + w_k] - w_j - \Theta\,, \quad j \in \mathbf{I},\ z \in \mathbf{Z}(j)\,,$$

$$Y_n = c(X_n, X_{n+1}; Z_n) - \Theta + w_{X_{n+1}} - w_{X_n} - \varphi(X_n, Z_n)\,, \quad n = 0, 1, \ldots$$

According to (1)

$$E\{Y_n \mid X_0, \ldots, X_n\} = 0\,, \quad n = 0, 1, \ldots$$

Hence,

$$(2) \qquad B_M = \sum_{n=0}^{M-1} Y_n = C_M - M\Theta + w_{X_M} - w_{X_0} - \sum_{n=0}^{M-1} \varphi(X_n, Z_n)\,, \quad M = 1, 2, \ldots,$$

is a martingale with respect to the observed state variables $\{X_0, X_1, \ldots, X_M\}$, $M = 1, 2, \ldots$ Using a system of equations widely employed in the theory of controlled Markov chains we shall determine the constants Θ and w_j, $j \in \mathbf{I}$, so that the correcting term will become asymptotically negligible. In this manner we obtain from the law of large numbers and from the central limit theorem for martingales the corresponding results for controlled Markov chain.

2. LAW OF LARGE NUMBERS

It holds

$$E\{Y_n^2 \mid X_0, \ldots, X_n\} = c_2(X_n, Z_n) - \varphi(X_n, Z_n)^2\,, \quad n = 0, 1, \ldots,$$

with

$$(3) \qquad c_2(j, z) = \sum_k p(j, k; z)\,[c(j, k; z) - \Theta + w_k - w_j]^2\,, \quad j \in \mathbf{I},\ z \in \mathbf{Z}(j)\,,$$

because of

$$E\{(c(X_n, X_{n+1}; Z_n) - \Theta + w_{X_{n+1}} - w_{X_n})\,\varphi(X_n, Z_n) \mid X_0, \ldots, X_n\} = \varphi(X_n, Z_n)^2\,.$$

The functions $c_2(j, z)$ and $\varphi(j, z)$ are bounded. Consequently,

$$\sum_{n=1}^{\infty} n^{-2} E Y_n^2 = \sum_{n=1}^{\infty} n^{-2} E(c_2(X_n, Z_n) - \varphi(X_n, Z_n)^2) < \infty\,.$$

The strong law of large numbers for martingales (see e.g. [4], § 29·1) then implies

$$0 = \lim_{M\to\infty} M^{-1}B_M = \lim_{M\to\infty} M^{-1}(C_M - M\Theta - \sum_{n=0}^{M-1}\varphi(X_n, Z_n)) \tag{4}$$

almost surely.

Consider a function $z(j)$ mapping j into $\boldsymbol{Z}(j), j \in \boldsymbol{I}$, with the following property:

Property 1. *The states $j \in \boldsymbol{I}$, which are recurrent for a Markov chain with transition matrix $\|p(j, k; z(j))\|_{j,k\in\boldsymbol{I}}$ form only one irreducible set.*

Property 1 implies that the constants Θ, w_j, $j \in \boldsymbol{I}$, can be chosen so that

$$\varphi(j, z(j)) = 0\,, \quad j \in \boldsymbol{I}\,. \tag{5}$$

(See e.g. Bellman's original paper [1] in a slightly less general setting.) We interpret the function $z(j)$, $j \in \boldsymbol{I}$, as a stationary control possessing certain desirable qualities and assume that the actual control variables $\{Z_n, n = 0, 1, \ldots\}$ approach this stationary control for $n \to \infty$.

Theorem 1. *Let* (5) *hold. If*

$$Z_n - z(X_n) \to 0\,, \quad \textit{for} \quad n \to \infty\,, \tag{6}$$

in probability [*almost surely*], *then*

$$\lim_{M\to\infty} M^{-1}C_M = \Theta \tag{7}$$

in probability [*almost surely*].

Proof. From the validity of (6) in probability, from (5) and from the continuity and boundedness of $\varphi(j, z)$ follows

$$\lim_{n\to\infty} E|\varphi(X_n, Z_n)| = \lim_{n\to\infty} E|\varphi(X_n, Z_n) - \varphi(X_n, z(X_n))| = 0\,.$$

Hence,

$$0 = \lim_{M\to\infty} M^{-1}\sum_{n=0}^{M-1} E|\varphi(X_n, Z_n)| \geqq \lim_{M\to\infty} E|M^{-1}\sum_{n=0}^{M-1}\varphi(X_n, Z_n)| \geqq 0 \tag{8}$$

(8) implies

$$\lim_{M\to\infty} M^{-1}\sum_{n=0}^{M-1}\varphi(X_n, Z_n) = 0 \tag{9}$$

in probability. From this and from (4) we get (7) in probability. Similarly, from (6) almost surely follows (9) almost surely and hence, (7) from (4).

We assume that (5) is fulfilled and that (6) holds in probability. We have

$$M^{-1}\sum_{n=0}^{M-1} E\{Y_n^2 \mid X_0, \ldots, X_n\} = M^{-1}\sum_{n=0}^{M-1} c_2(X_n, Z_n) - M^{-1}\sum_{n=0}^{M-1} \varphi(X_n, Z_n)^2 .$$

As in the proof of Theorem 1, (6) implies that the second term on the right-hand side tends to zero in probability. Moreover, according to Theorem 1,

$$\lim_{M\to\infty} M^{-1}\sum_{n=0}^{M-1} c_2(X_n, Z_n) = \sigma^2$$

in probability, where σ^2 is obtainable (together with auxiliary constants w_{2j}, $j \in \boldsymbol{I}$), from the equations

$$\sum_k p(j, k; z(j))[c_2(j, z(j)) + w_{2k}] - w_{2j} - \sigma^2 = 0 , \quad j \in \boldsymbol{I} . \tag{10}$$

(10) can be somewhat simplified. Namely, inserting $z(j)$ into (3) and using (5), we get

$$c_2(j, z(j)) = \sum_k p(j, k; z(j)) \left[(c(j, k; z(j)) - \Theta)^2 + \right.$$
$$\left. + 2(c(j, k; z(j)) - \Theta) w_k + w_k^2\right] - w_j^2, j \in \boldsymbol{I} .$$

Hence (10) is equivalent to

$$\sum_k p(j, k; z(j)) \left[(c(j, k; z(j)) - \Theta)^2 + (2(c(j, k; z(j)) - \Theta) w_k + w'_{2k}\right] - \tag{11}$$
$$- w'_{2j} - \sigma^2 = 0 , \quad j \in \boldsymbol{I} ,$$

where $w'_{2j} = w_{2j} + w_j^2$, $j \in \boldsymbol{I}$. (11) is the system of equations for the variance derived in [5].

Let us apply now the central limit theorem for martingales ([2], [3]). Suppose $\sigma^2 > 0$. Since the variables Y_n, $n = 0, 1, \ldots$, are bounded, the above established relation

$$\lim_{M\to\infty} M^{-1}\sum_{n=0}^{M-1} E\{Y_n^2 \mid X_0, \ldots, X_n\} = \sigma^2$$

in probability implies the validity of the central limit theorem, i.e. $B_M/\sqrt{M}$ has for $M \to \infty$ asymptotically normal distribution $N(0, \sigma^2)$. Thus, with regard to (2), we get the following result.

Theorem 2. *Let* (5) *and* (11) *hold with* $\sigma^2 > 0$. *If*

$$\lim_{n\to\infty} (Z_n - z(X_n)) = 0 = \lim_{M\to\infty} \sum_{n=0}^{M-1} \varphi(X_n, Z_n)/\sqrt{M}$$

in probability, then $(C_M - M\Theta)/\sigma\sqrt{M}$ *has for* $M \to \infty$ *asymptotically normal distribution* $N(0, 1)$.

A direct proof of a similar assertion was given in [6].

Remark. Suppose that Property 1 holds for all stationary controls. Then Θ, w_j, $j \in \boldsymbol{I}$, can be determined so that

$$\max_{z \in \boldsymbol{Z}(j)} \varphi(j, z) = 0\,, \quad j \in \boldsymbol{I}\,.$$

(Bellman's equations for the maximal mean reward [1].) From (4) we conclude that

$$\limsup_{M \to \infty} M^{-1} C_M = \Theta + \limsup_{M \to \infty} M^{-1} \sum_{n=0}^{M-1} \varphi(X_n, Z_n) \leqq \Theta\,,$$

almost surely for arbitrary $\{Z_n, n = 0, 1, \ldots\}$.

The paper was stimulated by the discussions which the author had with G. K. Eagleson in Cambridge.

(Received November 10, 1972)

REFERENCES

[1] R. Bellman: A Markovian decision process. J. of Math. and Mech. *6* (1957), 679—684.
[2] P. Billingsley: The Lindeberg-Lévy theorem for martingales. Proc. Amer. Math. Soc. *12* (1961), 788—792.
[3] B. M. Brown, G. K. Eagleson: Martingale convergence to infinitely divisible laws with finite variances. Trans. Amer. Math. Soc. *162* (1971), 449—453.
[4] M. Loéve: Probability theory. Princeton 1960.
[5] P. Mandl: On the variance in controlled Markov chains. Kybernetika *7* (1971), 1—12.
[6] P. Mandl: On the asymptotic normality of the reward in a controlled Markov chain. (To appear in Trans. of European Meeting of Statisticians held in Budapest, 1972.)

Dr. Petr Mandl, DrSc.; Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Vyšehradská 49, 128 48 Praha 2. Czechoslovakia.