E. DRABIK (Białystok)
L. STETTNER (Warszawa)

ON ADAPTIVE CONTROL OF MARKOV CHAINS USING NONPARAMETRIC ESTIMATION

Abstract. Two adaptive procedures for controlled Markov chains which are based on a nonparametric window estimation are shown.

1. Introduction. Assume on a probability space \((\Omega, F, P)\) we are given a discrete time controlled Markov process \(X = (x_i)\) with values in a finite state space \(E = \{1, \ldots, k\}\) and with an unknown transition matrix \(p^v(i, j)\) depending on a control parameter \(v \in [0, 1]\). Assume furthermore that for \(i, j \in E\) the mapping \([0, 1] \ni v \mapsto p^v(i, j)\) is continuous.

Our purpose is to minimize the following average cost per unit time functional:

\[
J_x(V) = \limsup_{n \to \infty} \frac{1}{n} E^V_x \left\{ \sum_{i=1}^{n} c(x_i, v_i) \right\}
\]

over all sequences \(V = (v_i)\) of \([0, 1]\)-valued \(\sigma\{x_0, \ldots, x_i\}\)-measurable random variables, where \(E^V_x\) stands for conditional expected value given that the controlled process \((x_i)\) starts from the state \(x\) and the control \(V\) is used, and \(c : E \times [0, 1] \to \mathbb{R}\) is a continuous function which measures the running cost.

An element \(u = [u_1, \ldots, u_k]\) of the set \(U = [0, 1]^k\) will be later interpreted as a Markov control in the sense that we shall use a control parameter equal to \(u_j\) when the state process \(x_i\) is in the state \(j\).

1991 Mathematics Subject Classification: 93E20, 93C40, 62M05.
Key words and phrases: adaptive control, controlled Markov chain, estimation.
The work was supported by KBN grant no. 2 P03A 01515.
Given a nondecreasing sequence \( \{b_n\}_{n \in \mathbb{N}} \) of positive integers such that \( b_n \to \infty \) as \( n \to \infty \), define the set
\[
\Phi(\{b_n\}_{n \in \mathbb{N}}) = \left\{ (a^n_i), \; i = 1, \ldots, n; \; n = 1, 2, \ldots; a^n_i \in \{0, 1\}, \; \sum_{i=1}^{n} a^n_i \geq b_n \right\}.
\]

The following auxiliary result will be used to justify the control procedures introduced in Sections 2 and 3.

**Proposition 1.** Let \( Y_i \) be a sequence of real-valued random variables such that \( E[Y_{i+1} \mid Y_1, \ldots, Y_i] = 0 \) and \( M = \sup_i E\{Y_i^2\} < \infty \). Then
\[
\sup_{(a^n_i) \in \Phi(\{b_n\})} \frac{\sum_{i=1}^{n} a^n_i Y_i}{\sum_{i=1}^{n} a^n_i} \to 0
\]
in probability as \( n \to \infty \).

**Proof.** Assume contrary to (2) that
\[
P\left( \frac{\sum_{i=1}^{n} a^n_i Y_i}{\sum_{i=1}^{n} a^n_i} \geq \varepsilon \right) \to 0
\]
does not converge to 0 as \( n \to \infty \) for \((a^n_i) \in \Phi(\{b_n\})\). Then by the Chebyshev inequality
\[
P\left( \frac{\sum_{i=1}^{n} a^n_i Y_i}{\sum_{i=1}^{n} a^n_i} \geq \varepsilon \right) \leq \frac{E[(\sum_{i=1}^{n} a^n_i Y_i)^2]}{\varepsilon^2 (\sum_{i=1}^{n} a^n_i)^2} \leq \frac{\sum_{i=1}^{n} (a^n_i)^2 (EY_i)^2}{\varepsilon^2 (\sum_{i=1}^{n} a^n_i)^2} \leq \frac{M}{b_n \varepsilon^2} \to 0
\]
and we have a contradiction. \( \blacksquare \)

Let \( \tilde{u}^i \in U, \; i = 0, 1, \ldots \), be a sequence equidistributed in \( U \). Given a sequence \( h_n \downarrow 0 \), for \( u = [u_1, \ldots, u_k] \in U \) and \( \tilde{u}^i_j \) being the \( j \)th coordinate of \( \tilde{u}^i \) define
\[
F_n(u) = \sum_{i=0}^{n} \prod_{j=1}^{k} 1_{\{|u_j - \tilde{u}^i_j| \leq h_n\}}.
\]
In what follows we shall assume that \( h_n \) is chosen such that
\[
f_n := \min_{u \in U} F_n(u) \to \infty
\]
as \( n \to \infty \). In particular, we can choose for \( \tilde{u}^i \) successively the centers of cubes with edges of length \( 1/2^j \) which cover the set \( U = [0, 1]^k \) for \( j = 0, 1, \ldots \) and consider the sequences \( \tilde{u}^i \) and \( h_i \) of the form
\[ \tilde{u}^0 = \left[ \frac{1}{2}, \ldots, \frac{1}{2} \right], \quad h_0 = \frac{1}{2}, \]
\[ \tilde{u}^1 = \left[ \frac{1}{4}, \frac{1}{4}, \ldots, \frac{1}{4} \right], \quad h_1 = \ldots = h_{2^k} = \frac{1}{3}, \]
\[ \tilde{u}^2 = \left[ \frac{3}{4}, \frac{1}{4}, \ldots, \frac{1}{4} \right], \]
\[ \vdots \]
\[ \tilde{u}^{2^k} = \left[ \frac{3}{4}, \frac{3}{4}, \ldots, \frac{3}{4} \right], \quad h_{2^k+1} = \ldots = h_{2^k+4^k} = \frac{1}{3}, \]
\[ \tilde{u}^{2^k+1} = \left[ \frac{1}{8}, \frac{1}{8}, \ldots, \frac{1}{8} \right], \]
\[ \vdots \]
\[ \tilde{u}^{2^k+4^k} = \left[ \frac{7}{8}, \frac{7}{8}, \ldots, \frac{7}{8} \right], \]
\[ \tilde{u}^{2^k+4^k+1} = \left[ \frac{1}{16}, \frac{1}{16}, \ldots, \frac{1}{16} \right], \quad h_{2^k+4^k+1} = \ldots = h_{2^k+4^k+8^k} = \frac{1}{5}, \]
\[ \tilde{u}^{2^k+4^k+8^k} = \left[ \frac{15}{16}, \frac{15}{16}, \ldots, \frac{15}{16} \right], \]
and so on.

In the theory of adaptive control of Markov processes the number of feasible procedures is very limited (see the papers [2], [4] and [5]). A recursive self-tuning algorithm proposed in [2] is based on asymptotic properties of ordinary differential equations and requires differentiability of the transition operator with respect to an unknown parameter. Other methods, in which discretized MLE ([4]) or the theory of large deviation ([5]) are used, require the construction of a finite class of \( \varepsilon \)-optimal controls, which is usually a hard problem.

In this paper we propose an alternative approach based on a window nonparametric estimation used for multiarmed bandit problems in [1]. Assuming that our model is uniformly ergodic (assumptions (5) and (13)), although we do not know the transition probabilities, it appears that we are able to construct an adaptive procedure for which we obtain self-optimality.

The paper consists of three sections. In Section 2 we introduce an adaptive procedure based on nonparametric estimation of the cost functional. In Section 3 another procedure that is based on nonparametric estimation of the transition kernel is considered.

2. Adaptive control with cost estimation. Assume there exists a uniformly positive recurrent state \( e \in E \) of the Markov process \( X \) in the sense that
\[
\sup_{u \in U} E^u_e \{ \tau^2 \} < \infty, \tag{5}
\]
where
\[
\tau = \inf\{ i > 0 : x_i = e \}. \tag{6}
\]
Note that the above property holds in particular when
\[ \inf_{v \in [0,1]} \inf_{j \in E} p^v(j,e) > 0. \]

Let
\[ \tau_1 = \tau, \ldots, \tau_{n+1} = \tau_n + \tau \cdot \Theta \tau_n \]
with \( \tau \) defined in (6) and \( \Theta \tau \) being the Markov shift operator. In other words \( \tau_n \) are the moments of successive returns to the recurrent state \( e \).

Assume now that in the time interval \( [\tau_i, \tau_{i+1}) \) we use a Markov control \( \tilde{u}^i \). For \( u \in U \) define
\[ G_n(u) = \sum_{i=0}^{n} \prod_{j=1}^{k} 1_{\{|u_j - \tilde{u}^i_j| \leq h_n\}} \sum_{r=\tau_i}^{\tau_{i+1}-1} c(x_r, \tilde{u}^i(x_r)) \]
and
\[ H_n(u) = \sum_{i=0}^{n} \prod_{j=1}^{k} 1_{\{|u_j - \tilde{u}^i_j| \leq h_n\}} (\tau_{i+1} - \tau_i). \]

Notice that \( G_n(u) \) is the total cost incurred in the time interval \( [0, \tau_{n+1}) \), when the control \( \tilde{u}^i \) from the closed ball with center \( u \) and radius \( h_n \) is used. Similarly \( H_n(u) \) is the total time during which the control from the sequence \( (\tilde{u}^i) \) lies in the closed ball with center \( u \) and radius \( h_n \).

**Proposition 2.** We have
\[ \sup_{u \in U} \left| \frac{G_n(u)}{F_n(u)} - E^u \left\{ \sum_{i=0}^{\tau-1} c(x_i, u(x_i)) \right\} \right| \to 0 \]
and
\[ \sup_{u \in U} \left| \frac{H_n(u)}{F_n(u)} - E^u \{ \tau \} \right| \to 0 \]
in probability as \( n \to \infty \), and consequently
\[ \sup_{u \in U} \left| \frac{G_n(u)}{H_n(u)} - \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \right| \to 0 \]
in probability as \( n \to \infty \), where \( \pi^u \) is the unique invariant measure corresponding to the Markov process \( X \) with Markov control \( u \).

**Proof.** Let
\[ Y_i = \sum_{r=\tau_i-1}^{\tau_i-1} c(x_r, \tilde{u}^i(x_r)) - E^{\tilde{u}^i} \left\{ \sum_{r=0}^{\tau_i-1} c(x_r, \tilde{u}^i(x_r)) \right\}, \]
with \( \tau_0 = 0 \). Clearly \( E[Y_i \mid Y_1, \ldots, Y_i] = 0 \) and from the boundedness of \( c(\cdot, \cdot) \) and (5) we have \( \sup_i E_i Y_i^2 < \infty \).
Consequently, from Proposition 1,
\[
\left| \sum_{i=0}^{n} \prod_{j=1}^{k} 1_{\{|u_j - \tilde{u}_j| \leq h_n\}} \left( \sum_{r=\tau_i+1}^{\tau_{i+1}-1} c(x_r, \tilde{u}_i(x_r)) - E^\nu_\tau \left( \sum_{r=0}^{\tau_{i+1}-1} c(x_r, \tilde{u}_i(x_r)) \right) \right) \right| \to 0
\]
converges to 0 in probability as \( n \to \infty \), uniformly in \( u \in U \).

Note that under (5), by continuity of \( p^v(e, j) \) with respect to \( v \), the mapping
\[
U \ni u \mapsto E^\nu \left\{ \sum_{r=0}^{\tau_i} c(x_r, u(x_r)) \right\},
\]
where \( U \) is endowed with the Euclidean norm, is continuous. Therefore, since \( h_n \to 0 \), we obtain
\[
\sup_{u \in U} \left| \frac{G_n(u)}{H_n(u)} - E^\nu_\tau \left\{ \sum_{r=0}^{\tau_i} c(x_r, u(x_r)) \right\} \right| \to 0
\]
in probability, which completes the proof of (10). The proof of (11) is similar.

We are now in a position to formulate our first control procedure:

For a given \( \varepsilon > 0 \) find a positive integer \( n_\varepsilon \) such that
\[
P \left\{ \sup_{u \in U} \left| \frac{G_n(u)}{H_n(u)} - \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \right| \geq \varepsilon \right\} \leq \frac{\varepsilon}{\|c\|}
\]
for \( n \geq n_\varepsilon \) with \( \| \cdot \| \) standing for the supremum norm.

For the first \( n_\varepsilon \) cycles, i.e. until time \( \tau_{n_\varepsilon+1} \), test controls from the sequence \( \tilde{u}_i \) using the Markov controls \( \hat{u}_i \) in the time intervals \((\tau_i, \tau_{i+1})\). At time \( \tau_{n_\varepsilon+1} \) find a control \( u_\delta \) that is \( \delta \)-optimal for \( G_{n_\varepsilon}(u)/H_{n_\varepsilon}(u) \), i.e. such that
\[
\frac{G_{n_\varepsilon}(u_\delta)}{H_{n_\varepsilon}(u_\delta)} \leq \inf_{u \in U} \frac{G_{n_\varepsilon}(u)}{H_{n_\varepsilon}(u)} + \delta,
\]
and use this control function for each \( i \geq \tau_{n_\varepsilon+1} \). Denote the above control procedure by \( V_c \).
Theorem 1. We have
\[ J_x(V_c) \leq \inf_{u \in U} \left[ \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \right] + 3\varepsilon + \delta. \]

Proof. Because of the form of the cost functional (1) and boundedness of the cost function \( c \), only controls after time \( \tau_{n+1} \) have an effect on the value of the cost functional \( J \) and therefore
\[ J_x(V_c) = E_x \left[ \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \right]. \]

Let
\[ B = \left\{ \sup_{u \in U} \left| \frac{G_{n_x}(u)}{H_{n_x}(u)} - \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \right| < \varepsilon \right\}. \]

For \( \omega \in B \) have
\[ \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \leq \varepsilon + \frac{G_{n_x}(u)}{H_{n_x}(u)} \leq \varepsilon + \delta + \inf_{u \in U} \frac{G_{n_x}(u)}{H_{n_x}(u)} \]
\[ \leq \inf_{u \in U} \left[ \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \right] + 2\varepsilon + \delta. \]

Consequently,
\[ J_x(V_c) \leq \inf_{u \in U} \left[ \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \right] + 2\varepsilon + \delta + \inf_{u \in U} P[1_{B^c} \|c\|] \]
\[ \leq \inf_{u \in U} \left[ \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \right] + 3\varepsilon + \delta, \]

which completes the proof. ■

3. Adaptive control procedure with transition probability estimation. In this section we estimate the transition probability function \( p^u(i, j) \). Assume now that the Markov process \( X = (x_i) \) is controlled using the sequence \( \tilde{u}^i \) at time \( i \). For \( l, l' \in E \) let (cf. (3) and (8))
\[ G_{n_x}^{l,l'}(u) = \sum_{i=0}^{n} \prod_{j=1}^{k} I_{|u_j - \tilde{u}^i_j| \leq h_n} 1_{l}(x_i) 1_{l'}(x_{i+1}) \]
and
\[ F_{n}(u) = \sum_{i=0}^{n} \prod_{j=1}^{k} I_{|u_j - \tilde{u}^i_j| \leq h_n} 1_{l}(x_i). \]
Assume now that there is \( \kappa > 0 \) such that for \( l, l' \in E \),

\[
\inf_{v \in [0,1]} p^v(l, l') > \kappa.
\]

**Proposition 3.** We have

\[
\sup_{u \in U} \left| \frac{G_{n,l'}^{l'}(u)}{F_n(u)} - p(l, l', u_l) \right| \to 0
\]

in probability as \( n \to \infty \), where \( p(l, l', v) := p^v(l, l') \).

**Proof.** Let

\[
Y_i = 1_l(x_i)1_{l'}(x_{i+1}) - 1_l(x_i)p(l, l', \tilde{u}_i).
\]

Clearly \( E[Y_i \mid Y_1, \ldots, Y_{i-1}] = 0 \). Therefore by Proposition 1,

\[
\sup_{u \in U} \sum_{i=0}^n \prod_{j=1}^k 1_{\{|u_j - \tilde{u}_j| \leq h_n\}} Y_i \to 0
\]

in probability as \( n \to \infty \). Using Proposition 1 again with \( Y_i' = 1_l(x_{i+1}) - p(x_i, l, \tilde{u}_{x_i}) \) we obtain

\[
\sup_{u \in U} \sum_{i=0}^n \prod_{j=1}^k 1_{\{|u_j - \tilde{u}_j| \leq h_n\}} Y_i' \to 0.
\]

Therefore by (13) for \( \kappa > \varepsilon > 0 \) we have

\[
P\left( \inf_{u \in U} \frac{F_n(u)}{F_n(u)} > \kappa - \varepsilon \right) \leq P\left( \sup_{u \in U} \sum_{i=0}^n \prod_{j=1}^k 1_{\{|u_j - \tilde{u}_j| \leq h_n\}} Y_i' \geq \varepsilon \right) \to 0
\]

in probability as \( n \to \infty \), and from (14) we obtain

\[
\sup_{u \in U} \sum_{i=0}^n \prod_{j=1}^k 1_{\{|u_j - \tilde{u}_j| \leq h_n\}} Y_i \to 0
\]

in probability as \( n \to \infty \). Hence

\[
\sup_{u \in U} \left| \frac{G_{n,l'}^{l'}(u)}{F_n(u)} - \sum_{i=0}^n \prod_{j=1}^k 1_{\{|u_j - \tilde{u}_j| \leq h_n\}} 1_l(x_i)p(l, l', \tilde{u}_i) \right| \to 0
\]

in probability as \( n \to \infty \) and by continuity of \( p(l, l', v) \) with respect to \( v \) we finally obtain

\[
\sum_{i=0}^n \prod_{j=1}^k 1_{\{|u_j - \tilde{u}_j| \leq h_n\}} 1_l(x_i)p(l, l', \tilde{u}_i) \to p(l, l', u_l)
\]

uniformly in \( u \in U \) as \( n \to \infty \), which completes the proof. \( \blacksquare \)
Our second adaptive procedure consists of the following steps:

1. For a given \( \varepsilon > 0 \) find \( n_\varepsilon \) such that for \( n \geq n_\varepsilon \) and \( l, l' \in E \),

\[
P \left\{ \sup_{u \in U} \left| \frac{G_{n_\varepsilon}^{l,l'}(u)}{F_{n_\varepsilon}(u)} - p(l, l', u_l) \right| > \varepsilon \right\} \leq \frac{\varepsilon}{\|c\|^2}.
\]

2. At time \( n_\varepsilon \) normalize \( G_{n_\varepsilon}^{l,l'}(u)/F_{n_\varepsilon}(u) \), i.e. form a new transition matrix

\[
\tilde{p}(l, l', u) = \frac{G_{n_\varepsilon}^{l,l'}(u)}{\sum_{r=1}^{k} G_{n_\varepsilon}^{l,r}(u)}.
\]

3. Then find an invariant measure \( \pi^u_\varepsilon \) for the transition matrix \( \tilde{p}(l, l', u) \) and determine \( u_\delta \) such that

\[
\sum_{\eta} c(\eta, u_\delta(\eta)) \pi^u_\varepsilon(\eta) \leq \inf_{u \in U} \sum_{\eta} c(\eta, u_\delta(\eta)) \pi^u(\eta) + \delta.
\]

4. Starting from time \( n_\varepsilon \) use the control function \( u_\delta \).

Denote the above control procedure by \( V_p \).

**Theorem 2.** Under (13) we have

\[
J(V_p) \leq \inf_{u \in U} \left[ \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) \right] + \|c\| \frac{(1 + k)\varepsilon}{(1 - k\varepsilon)\kappa} + \delta + \varepsilon,
\]

where \( \pi^u \) is the unique invariant measure corresponding to \( p(l, l', u_l) \).

**Proof.** If for each \( l, l' \in E \),

\[
\sup_{u \in U} \left| \frac{G_{n_\varepsilon}^{l,l'}(u)}{F_{n_\varepsilon}(u)} - p(l, l', u_l) \right| \leq \varepsilon
\]

then we have

\[
|\tilde{p}(l, l', u) - p(l, l', u_l)| = \frac{F_{n_\varepsilon}(u)}{\sum_{r=1}^{k} G_{n_\varepsilon}^{l,r}(u)} \left| \frac{G_{n_\varepsilon}^{l,l'}(u)}{F_{n_\varepsilon}(u)} - p(l, l', u_l) \sum_{l'_{=1}}^{n} \frac{G_{n_\varepsilon}^{l,l'}(u)}{F_{n_\varepsilon}(u)} \right| \leq \left( \varepsilon + p(l, l', u_l) \right) \left| 1 - \sum_{l'_{=1}}^{n} \frac{G_{n_\varepsilon}^{l,l'}(u)}{F_{n_\varepsilon}(u)} \right| \frac{1}{1 - k\varepsilon} \leq (1 + k)\varepsilon \frac{1}{1 - k\varepsilon}.
\]

From Theorem and Corollary 2 of [7] under (13) we see that for \( l \in E \),

\[
\sup_{u \in U} |\pi^u_\varepsilon(l) - \pi^u(l)| \leq \frac{1}{2} \cdot \frac{(1 + k)\varepsilon}{(1 - k\varepsilon)\kappa}.
\]

Therefore for \( \omega \in \bigcap_{l=1}^{k} \bigcap_{l'=1}^{k} B_{ll'} \), where
we have
\[ \sum_{\eta \in E} c(\eta, u(\eta)) \pi^{u, \eta}_\varepsilon(\eta) \leq \inf_{u \in U} \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) + \|c\| \frac{(1 + k)\varepsilon}{(1 - k\varepsilon)\kappa} + \delta. \]

Consequently,
\[ J(V_p) = E \left[ \sum_{\eta \in E} c(\eta, u(\eta)) \pi^{u, \eta}_\varepsilon(\eta) \right] \leq \inf_{u \in U} \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) + \|c\| \frac{(1 + k)\varepsilon}{(1 - k\varepsilon)\kappa} + \delta \]
\[ + \|c\| P \left( \bigcap_{l=1}^k \bigcap_{l' = 1}^k B_{ll'} \right) \]
\[ \leq \inf_{u \in U} \sum_{\eta \in E} c(\eta, u(\eta)) \pi^u(\eta) + \|c\| \frac{(1 + k)\varepsilon}{(1 - k\varepsilon)\kappa} + \delta + \varepsilon, \]

which completes the proof. \( \blacksquare \)

**Remark.** The adaptive procedures introduced in Sections 2 and 3 allow one to determine a nearly optimal control in a finite time. Using forcing and an increasing decision horizon from Section 3 of [3] for both adaptive procedures it is possible to construct optimal adaptive strategies.

**References**


Ewa Drabik  
Faculty of Economics  
University of Bialystok  
Warszawska 63  
15-062 Bialystok, Poland

Lukasz Stettner  
Institute of Mathematics  
Polish Academy of Sciences  
Sniadeckich 8  
00-950 Warszawa, Poland  
E-mail: stettner@impan.gov.pl

and

Warsaw School of Management and Marketing

Received on 13.11.1998;  
revised version on 27.8.1999