KYBERNETIKA — VOLUME 27 (1991), NUMBER 1

ESTIMATION AND ADAPTIVE CONTROL OF SPAN-CONTRACTING MARKOV DECISION PROCESSES*

GERHARD HÜBNER

For undiscounted Markov decision processes with unknown parameters a policy with maximal average (expected) reward may be obtained adaptively if these parameters are estimated with increasing precision and simultaneously the relative value function for the true parameters is approximated by using the presently best knowledge. The paper presents successive approximation methods in case of multi-step span-contraction. This assumption weakens one-step span-contraction (corresponding to one-step scrambling) used in earlier papers.

1. INTRODUCTION

For many types of sequential decision processes there are some parameters (represented by ϑ, say) which are known only approximately. So during the process new information on these parameters is gained. Therefore usually at each stage of the process decisions are chosen optimally with respect to the presently best knowledge of ϑ. This procedure is known as the method of estimation and control (cf. [11, 12, 14, 15]). But as long as ϑ is known only roughly it may be inefficient to calculate optimal decisions on this basis. Thus there are some approximating methods, especially the so called non-stationary value iteration (cf. [1, 2, 4, 7, 13]). Both methods and others too may be described in the following unified way (see [10]): Use any sequence u_n approximating the true value function w^{ϑ} and apply to this sequence at each stage a one-step optimal reward operator $\mathbf{U}^{\vartheta_n}$. Then the maximizing actions form an optimal policy for the true parameter ϑ. (Of course, the estimated parameters ϑ_n must converge to the true ϑ.)

But unfortunately this result was derived under a very strong convergence condition, i.e. the one-step scrambling condition which is not valid for many real applications,

* Presented at the "Kolloquium über Mathematische Statistik im Rahmen der Wissenschaftlichen Kolloquien der Universität Hamburg und der Karls-Universität Prag", Hamburg, June 1989.

cf. [10], Rem. 2.4. We present here the proof of the same result under the condition that for each parameter ϑ the optimal reward operator $\mathbf{U}^\vartheta$ is ν-step span-contracting (for some ν). Even this condition may be weakened to hold only for applying $\mathbf{U}^\vartheta$ to the true value function w^ϑ and any u_n out of the tail of the approximating sequence (see the Remark at the end of the paper).

On the other hand to insure the ν-step span contraction it is in general insufficient to have a ν-step scrambling condition. But the stronger "scrambling type condition" introduced in [5] will do (see also [3], p. 55 and p. 128). In the following we summarize the model, the assumptions, and the main result of [10] together with the new conditions and then present the pertinent new proof.

2. NOTATIONS AND ASSUMPTIONS

The model consists of a countable state space I, feasible (non-empty) decision sets $A(i)$, a (topological) parameter space Θ, the transition probabilities $p^\vartheta_{ij}(a)$ and the one-stage (real) rewards $r^\vartheta_i(a)$. The aim is to obtain policies with maximum average reward for the true parameter ϑ.

For shortness we introduce the set $K := \{(i, a),\ i \in I,\ a \in A(i)\}$ of feasible state-action pairs, the (feasible) state-action histories up to stage n $h_n := (i_0, a_0, i_1, \ldots \ldots, a_{n-1}, i_n) \in H_n$ and the set of (deterministic) policies

$$\Pi := \{\pi = (f_0, f_1, \ldots), f_n\colon H_n \to \bigcup A(i), f_n(h_n) \in A(i_n),\ n \in \mathbb{N}_0\}\ .$$

For bounded functions $u\colon I \to \mathbb{R}$ we use

$$\mathbf{L}^\vartheta\, u(i, a) := r^\vartheta_i(a) + \sum_{j\in I} p^\vartheta_{ij}(a)\, u(j)\,, \quad (i, a) \in K\,,$$

and

$$\mathbf{U}^\vartheta\, u(i) := \sup_{a\in A(i)} \mathbf{L}^\vartheta\, u(i, a)\,, \quad i \in I\ .$$

$g\colon I \to \bigcup A(i)$ with $g(i) \in A(i)$ is an ε-maximizer of $\mathbf{L}^\vartheta u$ if $\mathbf{L}^\vartheta u(i, g(i)) \geqq \mathbf{U}^\vartheta\, u(i) - \varepsilon$, $i \in I$.

Finally we write (as usual) for $u\colon B \to \mathbb{R}$ $\sup u := \sup_{x\in B} u(x)$, $\inf u := \inf_{x\in B} u(x)$, $\operatorname{sp} u := \sup u - \inf u$ and $\|u\| := \sup_{x\in B} |u(x)|$. Note that $\operatorname{sp} u \leqq 2\|u\|$ and, if $u(x) = 0$ for some $x \in B$, $\|u\| \leqq \operatorname{sp} u$.

We shall use the following assumptions

(B) (Boundedness): For all $\vartheta \in \Theta$ the reward r^ϑ is bounded on K.

(C) (Continuity): For $\vartheta' \to \vartheta$

$$\Delta r(\vartheta', \vartheta) := \sup_{(i,a)\in K} |r^{\vartheta'}_i(a) - r^\vartheta_i(a)| \to 0$$

$$\Delta p(\vartheta', \vartheta) := \sup_{(i,a\in K)} \sum_{j\in I} |p^{\vartheta'}_{ij}(a) - p^\vartheta_{ij}(a)| \to 0\ .$$

(ST) (Scrambling type condition): For all $\vartheta \in \Theta$ there is a $\nu = \nu^\vartheta \in \mathbb{N}$ and a $\varrho^\vartheta > 0$ such that

$$\sum_{l\in I} \min\,[p^\vartheta_{il}(\pi_\nu), p^\vartheta_{jl}(\sigma_\nu)] \geqq \varrho^\vartheta$$

for all $i \neq j (\in I)$ and all ν-stage policies $\pi_\nu = (f_0, \ldots, f_{\nu-1}), \sigma_\nu = (g_0, \ldots, g_{\nu-1})$ where $p_{il}^\vartheta(\pi_\nu)$, $p_{jl}^\vartheta(\sigma_\nu)$ are the corresponding ν-step transition probabilities.

Related to (ST) are the conditions (MS) and (S):

(MS) (Multi-step scrambling condition): (ST) holds only for $\pi_\nu = \sigma_\nu = (f, f, \ldots, f)$ for all f.

(S) (Scrambling condition): (MS) (or equivalently (ST)) holds with $\nu^\vartheta = 1$ for all $\vartheta \in \Theta$.

Further assumptions related to (ST) (see Lemma 1 below) are (MSC) and (SC):

(MSC) (Multi-step span contraction): There is a $\nu = \nu^\vartheta$ and $\varrho^\vartheta > 0$ with

$$\operatorname{sp}[(\mathrm{U}^\vartheta)^\nu u - (\mathrm{U}^\vartheta)^\nu v] \leqq (1 - \varrho^\vartheta) \operatorname{sp}(u - v)$$

for any bounded $u: I \to \mathbb{R}, v: I \to \mathbb{R}$.

(SC) (Span contraction): (MSC) holds for $\nu^\vartheta = 1$ for all $\vartheta \in \Theta$.

Since we do not discuss estimation methods in this paper we use

(E) (Existence of a strongly consistent sequence of estimates): There is a sequence $(\hat{\vartheta}_0, \hat{\vartheta}_1, \ldots)$ with $\hat{\vartheta}_n: H_n \to \Theta$ and $\hat{\vartheta}_n(\tilde{h}_n) \to \vartheta$ P_π^ϑ-a.s. $(n \to \infty)$ for all $\vartheta \in \Theta$, $\pi \in \Pi$ where P_π^ϑ is the canonical probability measure induced by π (and some starting condition) and $\tilde{h}_n$ is the random n-stage history.

For examples of sequences $(\hat{\vartheta}_n)$ see e.g. [2] and [11]–[14]. In [10] (and other papers) assumption (ST) is replaced by (S). For relations of the assumptions (B), (C), and (S) to assumptions used elsewhere see [10], Remark 2.3.

The assumptions (ST), (MS), (S), (MSC) and (SC) themselves are related as follows.

Lemma 1 (cf. [5], Th. 5, and [9], Th. 2).

(a) Trivially (S) ⇒ (ST) ⇒ (MS) and (SC) ⇒ (MSC).

(b) (ST) ⇒ (MSC), especially (S) ⇒ (SC).

(c) If ϱ^ϑ is replaced by 0 conditions (SC) and (MSC) hold without further assumptions.

The optimal average gain k^ϑ and the relative value function $(w^\vartheta(i),\ i \in I)$ for a fixed (known) ϑ are obtained as a solution of the optimality equation (Lemma 2) or as a limit of a successive approximation (Lemma 3). Especially w^ϑ will play a central role in the main theorem.

Lemma 2 (cf. [6], Th. 2.1). From (B) and (MS) or (MSC) follows:

(UBS) For fixed $\vartheta \in \Theta$ and $i_0 \in I$ there is a **u**nique **b**ounded **s**olution $(w^\vartheta, k^\vartheta)$, $w^\vartheta: I \to \mathbb{R}$, $k^\vartheta \in \mathbb{R}$, of

$$\mathrm{U}^\vartheta w^\vartheta = w^\vartheta + k^\vartheta\,, \quad w^\vartheta(i_0) = 0$$

and for this solution k^ϑ is the maximal gain (for ϑ) independent of the starting distribution.

Lemma 3 (cf. [6], Sec. 3, [3], Sec. 5.4). Assume (B) and (MS) or (MSC). For any bounded $v_0^\vartheta: I \to \mathbb{R}$ define

$$v_n^\vartheta := (\mathrm{U}^\vartheta)^n v_0^\vartheta,\ n \in \mathbb{N}\,, \quad \bar{v}_n^\vartheta(i) := v_n^\vartheta(i) - v_n^\vartheta(i_0)\,, \quad i \in I, \quad n \in \mathbb{N}_0\,.$$

Then v_n^ϑ is bounded, $\bar{v}_n^\vartheta \to w^\vartheta$, and $\mathbf{U}^\vartheta\, v_n^\vartheta(i_0) \to k^\vartheta$ $(n \to \infty)$. If (MSC) holds then in addition (with $\lfloor \cdot \rfloor$ denoting the integer part of a real number)

$$\operatorname{sp}(v_n^\vartheta - w^\vartheta) \geqq (1 - \varrho^\vartheta)^{\lfloor n/\nu^\vartheta \rfloor} \operatorname{sp}(v_0^\vartheta - w^\vartheta) \quad (n \in \mathbb{N})\,.$$

To obtain an optimal policy the unknown function w^ϑ has to be approximated by a sequence of functions u_n depending in the estimates $\vartheta_0, \vartheta_1, \ldots, \vartheta_n$ thus far obtained. In the classical case of "estimation and control" (u_n) will be the sequence (w^{ϑ_n}) (see Th. 2(a)).

Definition. A sequence of functions $(u_n^{\vartheta(n)}(i), i \in I,\ \vartheta^{(n)} := (\vartheta_0, \ldots, \vartheta_n) \in \Theta^{n+1})$ is called an admissible approximation for $(w^\vartheta, \vartheta \in \Theta)$ if $\operatorname{sp}(u_n^{\vartheta(n)} - w^\vartheta) \to 0$ for any sequence $(\vartheta_n, n \in \mathbb{N}_0)$ with $\vartheta_n \to \vartheta$ for $n \to \infty$. Note that $\operatorname{sp}(w_n - w^\vartheta) \to 0$ if and only if $\|\bar{w}_n - w^\vartheta\| \to 0$ with $\bar{w}_n := w_n - w_n(i_0)$.

Now we are ready to formulate the main result:

Theorem 1 (cf. [10], Th. 3.1). Assume (B), (C), (E), (UBS) and
(1) $(u_n^{\vartheta(n)}, n \in \mathbb{N}_0)$ is an admissible approximation for $(w^\vartheta, \vartheta \in \Theta)$,
(2) $(\varepsilon_n, n \in \mathbb{N}_0)$ is a sequence of errors $\varepsilon_n \to 0$ $(n \to \infty)$,
(3) $g_n^{\vartheta(n)}$ is an ε_n-maximizer of $\mathbf{L}^{\vartheta_n} u_n^{\vartheta(n)}$, $\vartheta^{(n)} \in \Theta^{n+1}$, $n \in \mathbb{N}_0$.
Then the policy $\tilde{\pi} := (g_n^{\tilde{\vartheta}(n)}(\tilde{\imath}_n), n \in \mathbb{N}_0)$ is average reward optimal for the true parameter ϑ.

Note that in Theorem 1 the assumption (UBS) can be replaced by (MS) or (MSC), cf. Lemma 2.

Theorem 2 (cf. [10], Th. 3.2). Assume (B), (C), (MSC). Then $(u_n^{\vartheta(n)}, n \in \mathbb{N}_0)$ is an admissible approximation for $(w^\vartheta, \vartheta \in \Theta)$ in the following cases:
(a) $u_n^{\vartheta(n)} := w^{\vartheta_n}$, $\vartheta^{(n)} \in \Theta^{n+1}$, $n \in \mathbb{N}_0$.
(b) $u_0^{\vartheta(0)} := u_0$ (u_0 bounded), $u_n^{\vartheta(n)} := \mathbf{U}^{\vartheta_{n-1}}\, u_{n-1}^{\vartheta(n-1)}$, $\vartheta^{(n)} \in \Theta^{n+1}$, $\quad n \in \mathbb{N}$.

Case (b) yields the "non-stationary value iteration" introduced by Federgruen, Schweitzer [4] and Baranov [2] (see also [1, 7, 13]). The right hand term of (b) is already calculated in the preceding step of the algorithm when $g_{n-1}^{\vartheta(n-1)}$ is determined according to (3) of Theorem 1. So at each step the optimal reward operator $\mathbf{U}^{\vartheta_n}$ has to be applied only once. This corresponds to the usual "value iteration". For modifications of (b) see [10], Remarks 3.4 to 3.6.

3. PROOF OF THEOREM 2

The proof of case (a) of Theorem 2 differs only inessentially from that of [10], Theorem 3.2(a). So only case (b) is proved here.

At first we need two lemmas.

Lemma 4 (Disturbed contraction). For any non-negative sequences (a_n) and (b_n) with $b_n \to 0$ and $a_{n+\nu} \leqq c a_n + b_n$ for some $\nu \geqq 1$, $0 \leqq c < 1$ follows $a_n \to 0$ $(n \to \infty)$.

Proof. By induction $a_{n+k\nu} \leqq c^k a_n + \sum_{i=1}^{k} c^{i-1} b_{n+(k-i)\nu} \leqq c^k a_n + 1/(1-c) \sup_{m \geqq n} b_m$. Setting n large enough for the second term to be small and then increasing k we obtain $a_{n+k\nu} \to 0$ $(k \to \infty)$ for all n and hence $a_n \to 0$. □

Lemma 5 (cf. [8], Th. 6.8, [16], Th. 6.1(b)). Using the notions of Assumption (C) we have $\|\mathbf{U}^{\vartheta'}v - \mathbf{U}^{\vartheta}v\| \leqq \Delta r(\vartheta', \vartheta) + \frac{1}{2}\Delta p(\vartheta', \vartheta)\,\mathrm{sp}\, v$ and hence $\mathrm{sp}\,(\mathbf{U}^{\vartheta'}v - \mathbf{U}^{\vartheta}v) \leqq$ $\leqq 2\,\Delta r(\vartheta', \vartheta) + \Delta p(\vartheta', \vartheta)\,\mathrm{sp}\, v$.

Proof of Theorem 2(b). We shall use the assumptions (B), (C), (MSC). We have to show for a fixed sequence (ϑ_n) with $\vartheta_n \to \vartheta$

$$a_n := \mathrm{sp}\,(u_n^{\vartheta(n)} - w^{\vartheta}) \to 0 \quad (n \to \infty)\,.$$

By Lemma 4 it is sufficient to show (for $n \geqq n_0$)

$$(1) \qquad a_{n+\nu} \leqq c a_n + b_n \quad \text{with} \quad 0 \leqq c < 1 \quad \text{and} \quad b_n \to 0\,.$$

Using $u_n := u_n^{\vartheta(n)}$, $\mathbf{U}_n := \mathbf{U}^{\vartheta_n}$, $\mathbf{U} := \mathbf{U}^{\vartheta}$, and $\mathrm{sp}\,(\mathbf{U}w^{\vartheta} - w^{\vartheta}) = 0$ (cp. Lemma 2) we obtain

$$a_{n+\nu} = \mathrm{sp}\,(u_{n+\nu} - w^{\vartheta}) \leqq$$
$$\leqq \mathrm{sp}\,(\mathbf{U}_{n+\nu-1} \dots \mathbf{U}_n u_n - \mathbf{U}^{\nu} u_n) + \mathrm{sp}\,(\mathbf{U}^{\nu} u_n - \mathbf{U}^{\nu} w^{\vartheta}) =: (\mathrm{I}) + (\mathrm{II})$$

From (MSC) we have

$$(\mathrm{II}) \leqq (1 - \varrho^{\vartheta})\,\mathrm{sp}\,(u_n - w^{\vartheta}) = (1 - \varrho^{\vartheta})\, a_n\,.$$

By splitting up we obtain

$$(\mathrm{I}) \leqq \sum_{l=0}^{\nu-1} \mathrm{sp}\,(\mathbf{U}^{\nu-l-1}\mathbf{U}_{n+l}\, u_{n+l} - \mathbf{U}^{\nu-l-1}\,\mathbf{U}\, u_{n+l})$$

and hence by Lemma 1(c) and Lemma 5

$$(\mathrm{I}) \leqq \sum_{l=0}^{\nu-1} \mathrm{sp}\,(\mathbf{U}_{n+l}\, u_{n+l} - \mathbf{U}\, u_{n+l}) \leqq \sum_{l=0}^{\nu-1} [2\,\Delta r(\vartheta_{n+l}, \vartheta) + \Delta p(\vartheta_{n+l}, \vartheta)\,\mathrm{sp}\, u_{n+l}].$$

Therefore with $\mathrm{sp}\, u_{n+l} \leqq \mathrm{sp}\,(u_{n+l} - w^{\vartheta}) + \mathrm{sp}\, w^{\vartheta}$ (w^{ϑ} is bounded)

$$a_{n+\nu} \leqq \nu b_n' + c_n' \sum_{l=0}^{\nu-1} a_{n+l} + (1 - \varrho^{\vartheta})\, a_n$$

where

$$c_n' := \max\,(\Delta p(\vartheta_{n+l}, \vartheta),\, 0 \leqq l \leqq n - 1) \to 0$$
$$b_n' := 2 \max\,(\Delta r(\vartheta_{n+l}, \vartheta), 0 \leqq l \leqq n - 1) + c_n'\,\mathrm{sp}\, w^{\vartheta} \to 0 \quad (n \to \infty)\,.$$

Analogously with (MSC) replaced by Lemma 1(c) for term (II) we have

$$a_{n+l} \leqq l b_n' + c_n' \sum_{m=0}^{l-1} a_{n+m} + a_n\,, \quad 1 \leqq l \leqq \nu - 1\,.$$

Straightforward (recursive) calculations yield

$$a_{n+\nu} \leqq b_n' \sum_{m=0}^{\nu-1} (c_n' + 1)^m + (c_n' + 1)^{\nu}\, a_n - \varrho^{\vartheta} a_n$$

(Iteration of the simpler one-step inequality $a_{n+1} \leqq b'_n + (c'_n + 1)\, a_n$ results in the same formula without the essential term $-\varrho^\vartheta a_n$, cp. the proof of Lemma 4.)

Now, if n is large enough we have $(c'_n + 1)^\nu - \varrho^\vartheta \leqq c < 1$ (for some c) and therefore (1) holds. □

Remark. As is to be seen from the preceding proof the condition (MSC) may be slightly weakened: The contraction inequality is needed only for $v = w^\vartheta$. In the same way the assumption (ST) is needed only for $\sigma_\nu = (f^\vartheta, \ldots, f^\vartheta)$ – if an optimal stationary policy $(f^\vartheta, f^\vartheta, \ldots)$ exists for each ϑ.

(Received December 8, 1989.)

REFERENCES

[1] R. S. Acosta-Abreu and O. Hernandez-Lerma: Iterative adaptive control of denumerable state average-cost Markov systems. Control Cybernet. *14* (1985), 313–322.

[2] V. V. Baranov: Recursive algorithms of adaptive control in stochastic systems. Cybernetics *17* (1981), 815–824.

[3] A. Federgruen: Markovian Control Problems. Math. Centre Tracts 97, Amsterdam 1983.

[4] A. Federgruen and P. J. Schweitzer: Nonstationary Markov decision problems with converging parameters. J. Optim. Theory Appl. *34* (1981), 207–241.

[5] A. Federgruen, P. J. Schweitzer and H. C. Tijms: Contraction mappings underlying undiscounted Markov decision problems. J. Math. Anal. Appl. *65* (1978), 711–730.

[6] A. Federgruen and H. C. Tijms: The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms. J. Appl. Probab. *15* (1978), 356–373.

[7] O. Hernandez-Lerma: Adaptive Control Processes. Springer-Verlag, Berlin—Heidelberg—New York 1989.

[8] K. Hinderer: On approximate solutions of finite-stage dynamic programs. In: Dynamic Programming and its applications (M. L. Puterman, ed.), Academic Press, New York 1978, pp. 289–317.

[9] G. Hübner: Contraction properties of Markov decision models with applications to the elimination of non-optimal actions. In: Dynamische Optimierung, Bonner Math. Schriften *98* (1977), 57–65.

[10] G. Hübner: A unified approach to adaptive control of average reward Markov decision processes. OR Spektrum *10* (1988), 161–166.

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

[12] M. Kurano: Adaptive policies in Markov decision processes with uncertain matrices. J. Inf. Optim. *4* (1983), 21–40.

[13] M. Kurano: Learning algorithms for Markov decision processes. J. Appl. Probab. *24* (1987), 270–276.

[14] P. Mandl: Estimation and control of Markov chains. Adv. in Appl. Probab. *6* (1974), 40–60.

[15] P. Mandl: On the adaptive control of countable Markov chains. In: Probability Theory, Banach Centre Publications, Warsaw 1979, pp. 159–173.

[16] W. Whitt: Approximations of dynamic programs. Math. Oper. Res. *3* (1978), 231–243.

Prof. Dr. Gerhard Hübner, Universität Hamburg, Institut für Mathematische Stochastik, Bundesstrasse 55, D-2000 Hamburg. Federal Republic of Germany.