A Unified Approach to Adaptive Control of Average Reward Markov Decision Processes

G. Hübner

Institut für Mathematische Stochastik, Universität Hamburg, Bundesstraße 55, D-2000 Hamburg 13

Received July 6, 1987 / Accepted in revised form January 25, 1988

Summary. The classical procedure for the adaptive control of average reward Markov decision processes with an unknown parameter chooses at each stage a decision which is optimal for the average reward problem with the presently estimated parameter. But in many cases it is inefficient or impossible to compute each time the long run optimal policy. So successive approximation methods were proposed and investigated. We present a unifying and generalizing approach including both types of methods mentioned above and generating a lot of new procedures, too.


1. Introduction

Many real world problems which may be described as stationary Markov decision models contain unknown parameters. About these parameters increasing information is obtained only when observing (and controlling) the ongoing process. A widely used average optimal procedure in this case, called the principle of estimation and control, is to act as if the presently best estimation would be the true parameter, cp. Kurano [10, 11], Mandl [13, 14]. Sometimes as e.g. for classical inventory problems the long run optimal policy is not too hard to be obtained, but in many cases it is inefficient or impossible to solve for each new parameter the whole infinite horizon problem. So successive approximation methods were proposed and investigated, cp. Baranov [2], Federgruen and Schweitzer [4], Acosta-Abreu and Hernandez-Lerma [1], and Kurano [12]. We present a unifying and generalizing approach including both types of methods: Assume that there is a sequence of estimators which converges for any policy almost certainly to the true parameter value (strong consistency) and a sequence of value functions each depending on the estimates thus far obtained and converging (span-converging) to the relative value function for the true parameter. Then a policy obtained by applying to these functions after each observation and estimation a policy improvement step (with the most recent parameter) is average reward optimal.

After stating definitions and assumptions (Sect. 2) we present this main theorem together with examples of the value functions used therein including the known procedures (Sect. 3). The proofs are postponed to Sect. 4.

2. Definitions and Assumptions

We consider a stationary Markov decision model with an additional unknown parameter where

- \( I \) is the denumerable state space,
- \( A(i) \) are the (non-empty) feasible decision sets \( i \in I \),
- \( K := \{(i, a), i \in I, a \in A(i)\} \) is the set of feasible state-action pairs,
- $\Theta$ is the (topological) parameter space
- $\tilde{p}_i^\vartheta$ is the probability of starting in $i \in I$ under $\vartheta \in \Theta$. $\tilde{p}_i^\vartheta$ may be concentrated on one state.
- $\tilde{p}_{ij}^\vartheta(a)$ is the (one step) transition probability from $i \in I$ to $j \in I$ under action $a \in A$ and parameter $\vartheta \in \Theta$.
- $\tilde{r}_i^\vartheta(i, a)$ is the (real) one stage expected reward in state $i \in I$ under $a \in A(i)$ and $\vartheta \in \Theta$.

In order to define policies and the pertinent values we use

$$ h_n := (i_0, a_0, i_1, \ldots, a_{n-1}) \in H_n := K^n \times I, $$

the state-action histories at stage $n \in \mathbb{N}_0$, and

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

the set of (deterministic) policies.

For any fixed policy $\pi \in \Pi$ and any $\vartheta \in \Theta$ in the usual way a probability measure $P_n^\vartheta$ on $K^{n_0}$ is defined. Let $\bar{h}_n = (\bar{i}_0, \bar{a}_0, \ldots, \bar{i}_n)$ be the random $n$-stage histories. Then we may define as usually the average expected reward (shortly the “gain”)

$$ \Psi_n^\vartheta := \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} \tilde{r}_{\bar{i}_n}^\vartheta(\bar{i}_n, \bar{a}_n) $$

and

$$ \Psi^\vartheta := \sup_{\pi \in \Pi} \Psi_n^\vartheta. $$

A policy $\pi$ with $\Psi_n^\vartheta = \Psi^\vartheta$ is average reward optimal for $\vartheta (\vartheta \in \Theta)$.

In order to assure bounded solutions for fixed and convergence of the approximating sequences we assume

Assumption A1. (Boundedness): For all $\vartheta \in \Theta$ the reward $\tilde{r}_i^\vartheta$ is bounded on $K$.

Assumption A2. (Scrambling Condition): For all $\vartheta \in \Theta$ there is a $\rho^\vartheta > 0$ such that

$$ \sum_{i \in I} \min_{a, a'} [\tilde{p}_{ij}^\vartheta(a), \tilde{p}_{ij}^\vartheta(a')] \geq \rho^\vartheta $$

for all $(i, a), (i', a') \in K, i \neq i'$.

Assumption A3. (Continuity): For $\vartheta' \rightarrow \vartheta$

$$ \sup_{(i, a) \in K} |\tilde{r}_{i}^\vartheta(i, a) - \tilde{r}_{i}^{\vartheta'}(i, a)| \rightarrow 0. $$

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

Assumption A4. There exists at least one sequence $(\tilde{\vartheta}_n, \tilde{\vartheta}_1, \ldots)$ of estimates $\tilde{\vartheta}_n : H_n \rightarrow \Theta$ which is strongly consistent, i.e.

$$ \tilde{\vartheta}_n(\tilde{h}_n) \rightarrow \vartheta \quad P^\vartheta - a.s. \quad (n \rightarrow \infty) \text{ for all } \vartheta \in \Theta, \pi \in \Pi. $$

For examples of strongly consistent estimates see e.g. Kurano [10–12], Mandl [13], Baranov [2].

Remark 2.1. We shall use e-optimal actions in our main theorem, so we do not need assumptions like e.g. the compactness of $A(i)$ and the continuity of $\tilde{r}_i^\vartheta(i, \cdot)$ and $\tilde{p}_{ij}^\vartheta$ on $A(i)$.

Remark 2.2. We shall restrict to deterministic policies, therefore we may omit measurability considerations since a deterministic policy may (essentially) be defined on the denumerable set of finite-length state histories, see Hinderer [7, p. 13].

Remark 2.3. The above assumptions contain as special cases the assumptions of Kurano [12] where state and action spaces are finite, transition matrices are strictly positive, and an estimation procedure is described explicitly.

They also contain the assumptions of Acosta-Abreu and Hernandez-Lerma [1] where $A1$ and $A2$ are stated uniformly in $\vartheta$, $A(i)$ has to be compact, and instead of $A3$ continuity of $r$ and $p$ in $a$ and $\vartheta$ is assumed. Strictly speaking, $A3$ is not implied by the assumptions of [1], but it is needed there in the proof. ($A3$ may be derived e.g. if additionally $UA(i)$ is assumed to be compact.)

The paper of Federgruen/Schweitzer [4] which assumes finite state and action spaces considers only one fixed sequence of parameters, and has (in the average reward case) much stronger continuity conditions than $A3$ which will in general not hold for the usual estimation schemes. (In fact their main purpose lies in other approximating concepts.) On the other hand they have no scrambling condition.

Remark 2.4. The scrambling Condition A2 is somewhat restrictive for practical applications; so it would be desirable to weaken it e.g. to a $\nu$-stage scrambling condition (cp. Federgruen/Tijms [3], Assumption C3). But in the present version Lemma 2.5(b) which follows from A2 is essential to the proof of Theorem 3.2(b) and the subsequent remarks. For Theorem 3.2(a), (c) (based on Lemma 2.6) weaker assumptions are sufficient.