Gerhard Hübner

Approximations for Markov decision problems

Acta Universitatis Carolinae. Mathematica et Physica, Vol. 24 (1983), No. 1, 35--40

Persistent URL: http://dml.cz/dmlcz/142503

1983 ACTA UNIVERSITATIS CAROLINAE – MATHEMATICA ET PHYSICA VOL. 24. NO. 1

Approximations for Markov decision problems

G. HÜBNER
Universität Hamburg, Institut für Mathematische Stochastik

Received 22 December 1982

Approximating methods in finite stationary Markov decision models are dealt with, viz., extrapolation bounds and sensitivity results applied to clustering of state and action spaces. Some remarks on structured models and on statistical problems are included.

Vyšetřují se přibližné metody v konečných stacionárních Markovových rozhodovacích modelech, zejména meze pro extrapolaci a citlivost řešení na agregaci v prostoru stavů či akcí. Několik poznámek je věnováno strukturálním a statistickým problémům.

Изучаются приближенные методы в конечных стационарных марковских моделях решений, именно оценки для экстраполяции и чувствительность результатов на агрегацию в пространствах состояний или действий. Несколько заметок посвящено структуральным и статистическим проблемам.

1. Introduction

Markov decision problems are those real world problems which may be modelled by a (mostly discrete time) Markov process controlled by decisions of (usually one) decision maker and endowed with a reward structure. The expected reward (total discounted or average) has to be maximized.

We shall give here some examples of this type form the real world:

- Inventory and production control:
  The level of inventory is influenced by ordering (and/or producing) decisions and by the random demand.
- Scheduling of jobs: The amount of work remaining to be done is influenced by allocating and sequencing decisions and by the random duration of jobs.
- Control of water resources: The level of water is influenced by the degree or release and the random inflow.

In the next section we shall introduce the basic concepts. Then the main approximating methods are treated, i.e., extrapolation bounds (Section 3) and sensitivity results applied to clustering of state and action spaces (Section 4). The final section is devoted to remarks on structured models and on statistical problems.

*) D-2000, Hambrug 13, Bundesstrasse 55, West Germany.

2. The model and fundamental methods

For clarity and simplicity we shall make some restrictions:

- We shall avoid models with more than one decision maker, the (so called) stochastic or Markov games.
- We shall use discrete time, although some semi-Markov problems may be included.
- We shall assume (except for the last section) that all data, i.e. all transition probabilities and rewards, are known.
- Finally we shall restrict to finite stationary models to avoid existence and measurability problems as well as additional indices.

A finite stationary *Markov decision model* consists of

- a finite or infinite number N of steps, called the horizon,
- a finite state space S containing all information from the past which is necessary for the future (to obtain a Markov process),
- finite sets D_s of actions available in state $s \in S$,
- a transition probability p where $p(s, a, s')$ is the probability to reach $s' \in S$ in one step when starting in $s \in S$ with action $a \in D_s$,
- a reward function r where $r(s, a)$ is the (possibly expected) reward for one step starting in $s \in S$ with action $a \in D_s$,
- a discounting factor $\beta > 0$ and
- in case of finite N a final reward $V^0(s)$ depending on the final state $s \in S$.

For this model we define a decision function f as a mapping which assigns to each $s \in S$ an action $a \in D_s$, shortly $f \in \mathop{\mathbf{X}}_{s \in S} D_s$.

A policy is composed of decision functions, according to the number of steps, i.e. $\pi = (f_0, f_1, \ldots)$; a stationary policy contains only identical decision functions, i.e. $f^\infty = (f, f, \ldots)$ for $N = \infty$. For each fixed policy π and any starting state s the *expected total discounted reward* is defined as

$$V_\pi(s) = E_{\pi s}\Big[\sum_{n=0}^{N-1} \beta^n r(X_n, f_n(X_n)) + \beta^N V^0(X_N)\Big]$$

where the probability measure depends on π and s, where X_n is the (random) state of the system at time point n (at the end of step n and the beginning of step $n + 1$), and where the last term is omitted if $N = \infty$. The *value function* V then is defined by

$$V(s) = \sup_\pi V_\pi(s)$$

and policy π^* is said to be $(\varepsilon -)$ *optimal*, if

$$V_{\pi^*}(s) \geqq V(s) \ (-\varepsilon) \quad \text{for all} \quad s \in S .$$

A *solution* of a Markov decision problem consists in deriving V *and* an optimal policy π^*.

The basic solution methods for the most important case $N = \infty$ and $\beta < 1$ are

based on the following two theorems (which may be found – also in more general versions – in every textbook on Markov decision problems, e.g. Howard (1960), Derman (1970)):

Theorem 1. V is the only solution of

$$v(s) = \sup_{a \in D_s} Lv(s, a) =: U\, v(s)$$

where

$$Lv(s, a) := r(s, a) + \beta \sum_{s'} p(s, a, s')\, v(s')$$

and f^∞ is optimal if $V(s) = LV(s, f(s)) =: L_f\, V(s)$ for all $s \in S$.

Theorem 2. $U^n V^0$ converges to V for $n \to \infty$ and for any V^0, where $U^n v$ is defined recursively by $U^0 v = v$, $U^n v = U(U^{n-1} v)$. The policy f^∞ is optimal if $U^n V^0 = L_f(U^{n-1} V^0)$ for an infinite number of n's.

In case of a fixed stationary policy f^∞ these theorems reduce to

Corollary 1. $V_f := V_{f^\infty}$ is the only solution of $v = L_f v$.

Corollary 2. $(L_f)^n\, V^0$ converges to V_f for $n \to \infty$ and for any V^0.

From these theorems two exact and some approximative solution methods are derived:

a. The system

$$v(s) \geqq Lv(s, a), \quad s \in S, \quad a \in D_s$$

$$\sum_s v(s) = \text{minimum!}$$

is solved by *linear programming* methods. By using the dual problem it is possible to obtain also an optimal stationary policy.

b. The *policy iteration* method works by the following steps:

(i) Choose an arbitrary V^0.

(ii) If V^n is given, calculate $UV^n = L_{f_n} V^n$.

(iii) If $V^n = UV^n$ then (V^n, f_n^∞) is a solution, if $V^n \neq UV^n$ calculate $V^{n+1} := := V_{f_n}$ by solving the linear system $v = L_{f_n} v$ (Corollary 1).

This procedure will end in a finite number of steps.

c. If in b(iii) V_{f_n} is calculated iteratively by Corollary 2 then policy iteration turns to be also an approximative method.

d. The *value iteration* method works according to Theorem 2 by starting with V^0 and iterating $V^n = UV^{n-1}$ "sufficiently often".

e. The *policy-value-iteration* (e.g. van Nunen (1976)) is a variant of the policy iteration by setting $V^{n+1} = (L_{f_n})^{\lambda_n} V^n$ where λ_n are positive integers (possibly independent of n). This variant does no longer end after finitly many steps. Only if

(formally) $\lambda_n \equiv \infty$ then the ordinary policy iteration results (cp. Corollary 2). On the other hand for $\lambda_n \equiv 1$ we obtain the value iteration method.

For the methods c, d and e terminating rules and accuracy bounds are neccessary which may be derived from the next section.

3. Extrapolation bounds

The first who gave useful upper and lower bounds for the value function was MacQueen (1966). These bounds read, adapted to value iteration ($V^n = UV^{n-1}$)

$$V^N - V^k \leqq \sum_{i=1}^{N-k} \beta^i \sup_{\inf} (V^k - V^{k-1})$$

where N may be finite ($> k$) or infinite.

These bounds turned out to be very good for a moderate size of k in contrast to the bounds used earlier-based on the supremum norm.

This fast convergence is related to the fact that for a fixed stationary policy f^∞ and in the frequent case of an irreducible and aperiodic transition matrix p_f ($p_f(s, s') = p(s, f(s), s')$) has a single largest eigenvalue 1 with a constant right eigenvector and all other eigenvalues are of an absolute value less than one.Therefore the difference $V^n - V^{n-1}$ approaches asymptotically $d \,.\, \beta^n$ where d is a constant vector. So the above bounds will tend to be close together (cp. e.g. Schellhaas (1974)).

In the case not treated here in detail where the matrices p_f do not have equal row sums (e.g. stopped decision processes or transformed semi-Markov problems) in the asymptotic expansion $d \,.\, \beta^n$ (see above) d is not longer constant and β is not longer known. So β and d have to be estimated by

$$\hat{\beta} \approx \frac{V^k - V^{k-1}}{V^{k-1} - V^{k-2}}, \quad \hat{d} \approx \hat{\beta}^{-k}(V^k - V^{k-1}).$$

In more detail these results may be found in Schellhaas (1974) for $N = \infty$ and in Hübner (1980) for finite N.

By application of the bounds described above it is possible, too, to eliminate some actions at early stages which will be non-optimal later on (see e.g. MacQueen (1967), Hastings/van Nunen (1977) and Hübner (1977, 1979)).

4. Sensitivity and clustering

First we ask for the impact of inaccurate data on the value function. By similar methods to those used in Section 3 the following bounds are obtained:
If V^0 is changed to $\tilde{V}^0$ and therefore V^n to $\tilde{V}^n$ then

$$V^n - \tilde{V}^n \leqq \sup_{\inf} (V^0 - \tilde{V}^0) \,.\, \beta^n \,.$$

Similarly, if r is changed to $\tilde{r}$ then

$$V^n - \tilde{V}^n \leqq \sup_{\inf}(r - \tilde{r}) \cdot \sum_{k=0}^{n-1} \beta^k .$$

For changing p to $\tilde{p}$ the result is

$$V^n - \tilde{V}^n \leqq \pm \sup_{s,a} \sum_j (p(s,a,j) - \tilde{p}(s,a,j))^+ \cdot \sum_{k=0}^{n-1} \beta^{n-k} \operatorname{sp} V^k$$

where $\operatorname{sp} V^k = \sup V^k - \inf V^k$. Finally by changing β to $\tilde{\beta}$ we obtain

$$V^n - \tilde{V}^n \leqq (\beta - \tilde{\beta}) \cdot \sum_{k=0}^{n-1} \beta^{n-k-1} \sup_{\inf} V^k .$$

When changing more than one entry these formulas may be combined in an appropriate order: make sure that V^k in the third or fourth inequality are the known ones and avoid if possible the unknown $\tilde{V}^k$ on the right hand side.

These sensitivity results may be used if the state and action spaces are too large, possibly infinite. Then some states or actions are clustered together and a common (intermediate) value is chosen for r or V^0 (e.g.) on each cluster. So the original functions are compared with step functions on the original state and action spaces and the above inequalities may be applied whereas policy value iteration is carried out in the reduced spaces. Such methods and inequalities may be found in Fox (1971), Bertsekas (1975), White (1977), Whitt (1978) and Hinderer (1978).

A somewhat different approach is used by Nollau and Hahnewald-Busch (1979) who calculate upper and lower bounds $\overline{V}^n$ and $\underline{V}_n$ in upper and lower clustered models.

5. Structural and statistical problems

The results given so far do not take advantage of the special structure the decision problem may posess. So these methods have to be adapted to the special structures to yield simpler calculations and even simple structured policies. For practical purposes it will even be better in most cases to have simple suboptimal policies than complicated optimal ones. Possibly the most famous result of this type is the optimality of fixed lower and upper ordering bounds (so called (s, S)-policies) for inventory problems. There is a lot of results on such structured problems but even more seem to be unsolved.

Finally we have to discuss the important fact that all data needed for a decision model have to be collected by statistical methods.

It is the *simplest way* first to collect and evaluate data and then use these in a Markov decision model. But proceeding in this way the data and insights gained during the process are not used to update the values previously determined.

This disadvantage may be circumvented by at least three ways:

(i) Choose a *Bayes model* including an apriori distribution of the unknown para-

meters. But thus many more data have to be used and even the state space has to be enlarged by the posterior distributions to retain a Markov model.

(ii) Solve the decision problem with the data collected initially, but use only one or a few steps of the policy determined, then *update* your data and solve the decision problem again. This method is used in some applications and there are investigations on the asymptotical behaviour of such a proceeding (cp. e.g. Mandl (1979)).

(iii) Use methods of *time series analysis* of filtering to combine statistics and optimization. But as far as I know only few has been done in this direction.

References

[1] BERTSEKAS, D. P.: (1975) Convergence of discretization procedures in dynamic programming. *IEEE Trans. Autom. Control 20,* 415—419.

[2] DERMAN, C.: (1970) Finite state Markovian decision processes. *Academic Press,* New York.

[3] FOX, B. L.: (1971) Finite-state approximations in denumerable-state dynamic programs. *J. Math. Anal. Appl. 34,* 665—670.

[4] HASTINGS, N. A. J./VAN NUNEN, J. A. E. E.: (1977) The action elimination algorithm for Markov decision processes. In Markov Decision Theory ed. H. C. Tijms and J. Wessels, *Mathematical Centre Tracts,* Amsterdam *93,* 161—170.

[5] HINDERER, K.: (1978) On approximate solutions of finite-stage dynamic programs. In Dynamic Programming and its Applications, ed. M. L. Puterman, *Academic Press,* New York, 289—317.

[6] HOWARD, R. A.: (1960) Dynamic Programming and Markov Processes. *Wiley,* New York.

[7] HÜBNER, G.: (1977) Improved procedures for elimination suboptimal actions in Markov programming by the use of contraction properties. *Trans. 7th Prague Conf.* 1974, Reidel, Dordrecht, *Vol. A,* 257—263.

[8] HÜBNER, G.: (1979) Sequential similarity transformations for solving finite-stage sub-Markov decision problems. *Methods of Oper. Res. 33,* 197—207.

[9] HÜBNER, G.: (1980) Bounds and good policies in stationary finitestage Markovian decision problems. *Adv. Appl. Prob. 12,* 154—173.

[10] MACQUEEN, J.: (1966) A modified dynamic programming method for Markovian decision problems. *J. Math. Anal. App. 14,* 38—43.

[11] MACQUEEN, J.: (1967) A test for suboptimal actions in Markovian decision problems. *Operat. Res. 15,* 559—561.

[12] MANDL, P.: (1979) On the adaptive control of countable Markov chains. In: *Probability theory, Banach Centre Publ. 5,* PWN ,Warsaw, 159—173.

[13] NOLLAU, V./HAHNEWALD-BUSCH, A.: (1979) An approximation procedure for stochastic dynamic programming based on clustering of state and action spaces. *Math. Operationsforschung Statist. Ser. Opt. 10,* 121—130.

[14] VAN NUNEN, J. A. E. E.: (1976) Contracting Markov Decision Processes. Doctoral Dissertation. University of Technology, Eindhoven.

[15] SCHELLHAAS, H.: (1974) On the extrapolation in Markov decision models with discounting (in German). *Z. Oper. Res. 18,* 91—104.

[16] WHITE, D. J.: (1977) Finite state approximations for denumerable state infinite horizon disconted Markov decision processes. Note Nr. 43, *Dept. of Decision Theory,* Univ. of Manchester.

[17] WHITT, W.: (1978) Approximations of dynamic programs I. *Math. Oper. Res. 3,* 231—243.