MODELING, IDENTIFICATION AND CONTROL, 1987, VOL. 8, NO. 2, 117–123
doi:10.4173/mic.1987.2.4

Finite-state approximations for countable-state infinite horizon discounted Markov decision processes

SJUR D. FLÅM†

Keywords: *Markov decision processes, approximation, epiconvergence.*

It is proved that the optimal policy of a Markov decision process where the state space is truncated, will approximate the policy in case of no truncation.

1. Introduction

Many finite horizon problems of economic control can be formulated as Markov Decision Processes (MDP). Moreover, for a discounted homogeneous MDP one may theoretically solve for an optimal strategy using the method of successive approximations, policy improvement, or linear programming (Bertsekas and Shreve 1978, Denardo 1987, Ross 1983).

In many practical cases, however, the computational burden becomes excessive and we are forced to take steps to ease it. One avenue is to implement acceleration techniques by including standard Gauss-Seidel procedures (Denardo 1982) or tests to eliminate sub-optimal strategies (MacQueen 1967).

Having done this, the well known curse of dimensionality is likely to remain a major deterrent for efficient computation. To mitigate this hindrance we usually have to content ourselves with good approximations, i.e. the original problem must be replaced by a simpler one. We have several options available. We may

discretize the state or action space (Bertsekas 1976),

concentrate the probability distributions governing the system dynamics (Flåm 1987, Norman and White 1968, Kall 1987),

decompose large programs into independent modules,

aggregate states, actions or revenues (Mendelssohn 1980),

or finally, we may ignore parts of the state space (Fox 1971, White 1980).

However, with any of these approximation schemes we face two problems:

If it is assumed that as the integer n increases towards infinity, the optimization problem ρ_n provides an improving approximation to the more difficult original problem ρ, then, we must first demonstrate that the optimal *values* of ρ_n converge to the optimal value of ρ. Secondly, and equally important, it must be shown that optimal *solutions* of ρ_n cluster to those of ρ in some sense as n moves towards infinity. For general information on this see (Kall 1986), and for results on MDP consult (Whitt 1978, 1979).

This paper is concerned with these two problems for the last scheme of approximation mentioned above, namely the procedure of truncating the state space. In the literature we have received on this procedure, we do not find any argument supporting the statement that optimal solutions of the approximate problems cluster to

Received 16 October 1986.

Institute of Economics, University of Bergen, Jonas Reinsgate 19, 5008 Bergen, Norway.

those of the original problem. We regard this statement to be crucial. Therefore the objective of this paper is to give a short proof for its validity.

The paper is organized as follows. Section 2 briefly recalls the MDP and the associated optimization problem. Section 3 introduces the scheme of approximation and furnishes the result concerning the convergence of optimal solutions.

2. Preliminaries

Let $x_0, x_1, x_2, \ldots$ denote the trajectory in the state space X of a discrete time stochastic process.

The sequence $(x_t)_{t=1}^{\infty}$ is controlled by a decision maker who at each time $t = 0, 1, \ldots$. after observing x_t chooses a control $u_t \in U$. When doing so a cost $c(x_t, u_t)$ is immediately incurred, and the system moves on to the next state x_{t+1} according to the probability distribution $P(x_t, x_{t+1}; u_t)$. The initial state x_0 has the distribution μ. Future costs are discounted by a constant factor $\delta \in (0, 1)$.The overall objective is to minimize the expected present value of the flow of costs. Gihman and Skorohod (1979) prove that for this minimization problem only stationary Markovian and non-randomized policies need be considered. Such a policy amounts to a rule $\pi: X \to U$ for picking a definite control $\pi(x)$ at any stage which only depends upon the current state.

The expected present value of such a policy is denoted by $v(\pi)$. Our goal is to give approximate solutions to the following problem

$$\rho\text{: Find } \pi^* \text{ such that } v(\pi^*) = \inf\{v(\pi) | \pi \in \Pi\}.$$

In order to simplify and to make problem ρ well defined we shall impose the assumption in 2.1.

2.1. Assumptions

The state space X is countable.
The control space U is compact.
The transition probability preserves continuity in the sense that for any bounded f: $X \to R$, the expectation

$$\sum_y f(y)P(x, y; u)$$

depends continuously on u.
The cost function $c(x, u)$ is lower semi-continuous in u and bounded.

Under these assumptions which will be used throughout this paper, the following theorem is a consequence of a more general result in (Gihman and Skorohod 1979, Theorem 1.13).

2.2 Theorem

Under assumptions 2.1 an optimal policy π^*: $X \to U$ exists where $\pi^*(x)$ is obtainable from the equation

$$V(x) = c(x, \pi^*(x)) + \delta \sum_y V(y)P(x, y; \pi^*(x)).$$

Here the optimal value function $V: X \to R$ is the unique solution of the functional equation

$$V(x) = \inf_u \left[c(x, u) + \delta \sum_y V(y)P(x, y; u) \right]$$

Moreover, $v(\pi^*) = \sum_x V(x)\mu(x)$.

3. **The scheme of approximation** (Fox 1971), (White 1980).

Let $X_1 \subset X_2 \subset \ldots \subset X$ be a tower of finite subsets of the state space X

so that
$$\overset{\infty}{\underset{n=1}{U}} X_n = X.$$

For each $n \geq 1$, define the data of an approximate MDP$_n$ in the following fashion: Let

$$c_n(x, u) = \begin{cases} c(x, u) & \text{if } \quad x \in X_n \\ 0 & \text{otherwise} \end{cases}$$

Furthermore, let $P_n(x, y; u) = P(x, y; u)$ if $x \in X_n$, otherwise define $P_n(x, y; u)$ to equal the unit measure concentrated at x.

Thus instead of controlling the original MDP having data (μ, δ, c, P) we propose to control another approximate MDP$_n$ with data (μ, δ, c_n, P_n). $v_n(\pi)$ denotes the value function of this approximate MDP$_n$ when we implement the policy π. We now agree that

$$\pi_n^* \in \operatorname{argmin} v_n$$

denotes the fact that π_n^* is an optimal solution to the following minimization problem:

$$\rho_n\text{: Find } \pi_n^* \text{ such that } v_n(\pi_n^*) = \inf_\pi v_n(\pi)$$

The expression

$$\pi^* \in \operatorname{argmin} v$$

is similarly defined with reference to the original problem ρ.

According to Theorem 2.2., argmin v_n, $n \geq 1$ and argmin v are nonempty sets. In order to obtain the results on the convergence results we will impose an additional assumption, 3.1.

3.1. Assumption

The control space U is metrizable.

Recall that U is already required to be compact. Hence it is also separable. We are now in a position to state the chief result of this paper.

3.2. Theorem

Assumptions 2.1 and 3.1 imply that:

(i) *If* $\pi^*_{n_k} \in \operatorname{argmin} v_{n_k} \to \pi^*$ *pointwise for some subsequence* n_k, *then* $\pi^* \in \operatorname{argmin} v$, *and moreover*, $\inf_\pi v_{n_k}(\pi) \to \inf_\pi v(\pi)$ *as* $k \to \infty$.

(ii) *Given any sequence* $\pi_n \in \operatorname{argmin} v_n$, $n \geqslant 1$, *we may then extract a subsequence* π_{n_k}, $k = 1$,. *which converges pointwise.*

The proof of statements (i) and (ii) is organized so as to apply results from the theory of epi-convergence (Attouch and Wets 1981, 1983).

3.3. Definition

The sequence $f_n\colon \Pi \to [-\infty, \infty]$, $n \geqslant 1$ *of extended real-valued functions defined on a metric space* Π *is said to epi-converge to* $f\colon \Pi \to [-\infty, \infty]$ *at* π *if*

(*a*) *for any subsequence* f_{n_k}, $k = 1, 2, \ldots$ *and any sequence* $\pi_k \in \Pi$ *converging to* π, $\lim_{k\to\infty} \inf f_{n_k}(\pi_k \geqslant f(\pi)$, *and*

(*b*) *there exist a sequence* π_n, $n = 1, 2, \ldots$ *converging to* π *such that* $\lim_{n\to\infty} \sup f_n(\pi_n) \leqslant f(\pi)$

If (*a*) and (*b*) are satisfied for every $\pi \in \Pi$, we say that f is the *epi-limit* of the sequence f_n, $n \geqslant 1$ or, alternatively, that f_n *epi-converges* to f.

This concept, which emerged from the study of approximation schemes, is useful as can be seen by the following statement.

3.4. Theorem (Attouch and Wets 1983)

Suppose f *is the epi-limit of* f_n, $n \geqslant 1$, *and* $\pi^*_{n_k} \in \operatorname{argmin} f_{n_k}$, $\pi^*_{n_k} \to \pi^*$. *Then* $\pi^* \in \operatorname{argmin} f$ *and* $\lim_{k\to\infty} (\inf f_{n_k}) = \inf f$.

In our setting let $\Pi = U^X$ be the set of all functions $\pi\colon X \to U$. Endow Π with the product topology. The compactness of U implies that Π is also compact. Moreover, since X is countable and U is metrizable, Π is also metrizable.

If the sequence of the value functions v_n, $n \geqslant 1$ epi-convergences to v, then Theorem 3.4 immediately justifies statement (i) of Theorem 3.2. Thus it only remains to prove the following auxiliary result.

3.5. Theorem

Under the assumptions in 2.1 and 3.1 the value functions v_n, $n \geqslant 1$, *epi-convergences to* v.

Proof: Use l_∞ to denote the Banach space of all bounded functions $f\colon X \to R$ under the supremum norm. For each $n \geqslant 1$ and each policy π, define the operator $A_n(\pi)$ on l_∞ in the following way:

$$A_n(\pi)f(x) = c_n(x, \pi(x)) + \delta \sum_y f(y) P_n(x, y; \pi(x))$$

The operator $A(\pi)$ is defined similarly. It is straightforward to show that $A_n(\pi)$, $n \geqslant 1$ and $A(\pi)$ are contraction operators on l_∞. Moreover, the fixed points $V_n(\pi)$ and $V(\pi)$, respectively satisfy $v_n(\pi) = E_\mu V_n(\pi)$ and $v(\pi) = E_\mu V(\pi)$ where E_μ denotes the expectation with respect to the initial distribution μ.

Thus $v_n(\pi) = E_\mu \lim_{k\to\infty} A_n^k(\pi)f$ for any $f \in l_\infty$. The formula for $A_n(\pi)$ can be simplified. If we assume that f equals zero outside X_n, then $A_n(\pi)f$ evidently also vanishes outside X_n. Consequently we may redefine $A_n(\pi)f(x)$ to be equal to

$$c(x, \pi(x)) + \delta \sum_{y \in X_n} f(y)P(x, y; \pi(x)) \tag{*}$$

on X_n and to equal zero otherwise.

Since the cost function $c(x, u)$ is bounded we may assume, without loss of generality, that $c(x, u) \geqslant 0$ for all $x \in X$, $u \in U$.

If (*) holds, it is evident that

$$A_n(\pi)f(x) \leqslant A_{n+1}(\pi)g(x) \leqslant A(\pi)h(x)$$

for all $x \in X$ whenever $f, g, h \in l_\infty$ satisfy $0 \leqslant f(x) \leqslant g(x) \leqslant h(x)$ for all $x \in X$. Hence

$$\begin{aligned} V_n(\pi)(x) &= \lim_{k\to\infty} A_n^k(\pi)0(x) \leqslant \lim_{k\to\infty} A_{n+1}^k(\pi)0(x) \\ &= V_{n+1}(\pi)(x) \leqslant \lim_{k\to\infty} A^k(\pi)0(x) = V(\pi)(x) \end{aligned}$$

where $0 \in l_\infty$ is defined by $0(x) \equiv 0$.

It follows that for all $\pi \in \Pi$, $v_n(\pi) = \mathrm{E}_\mu V_n(\pi)$ is an increasing sequence bounded from above by $v(\pi)$. In particular this takes care of condition (*b*) in Definition 3.3.

We now turn to condition (*a*) of that definition. Suppose $\pi_k \to \pi$ and let v_{n_k}, $k = 1, \ldots$ be a subsequence of v_n, $n \geqslant 1$. For arbitrary $\varepsilon > 0$ choose an integer M such that $m \geqslant M$ implies

$$E_\mu A^m(\pi) \geqslant v(\pi) - \varepsilon.$$

Note that for any integer n, policy π and state x, the sequence $A_n^m(\pi)0(x)$, $m \geqslant 1$ is non-decreasing. Hence

$$\begin{aligned} \liminf_{k\to\infty} v_{n_k}(\pi_k) &= \liminf_{k\to\infty} E_\mu \lim_{m\to\infty} A_{nk}^m(\pi_k)0 \\ &\geqslant \liminf_{k\to\infty} E_\mu A_{nk}^M(\pi_k)0 \\ &\geqslant E_\mu \liminf_{k\to\infty} A_{nk}^M(\pi_k)0 \\ &\geqslant E_\mu A^M(\pi)0 \geqslant v(\pi) - \varepsilon \end{aligned}$$

For the second inequality we have applied Fatou's lemma, for the third we needed the lower semi-continuity of c and lemma 1.5 of (Gihman and Skorohod 1979).

Specifically, let by convention $A^0(\pi)0(x) = 0$ for all x, and suppose inductively that

$$\liminf_{k\to\infty} A_{nk}^t(\pi_k)0(x) \geqslant A^t(\pi)0(x)$$

for some integer $t \geqslant 0$, and all x. Then

$$\begin{aligned}\liminf_{k\to\infty} A^{t+1}_{n_k}(\pi_k)0(x) &\geqslant \liminf_{k\to\infty}\Big[c_{n_k}(x, \pi_k(x)) \\ &\qquad + \delta \sum_y A^t_{n_k}(\pi_k)0(y)P_{n_k}(x, y; \pi_k(x))\Big] \\ &\geqslant c(x, \pi(x)) + \delta \sum_y A^t(\pi)0(y)P(x, y; \pi(x)) \\ &= A^{t+1}(\pi)0(x).\end{aligned}$$

This takes care of the crucial inequality

$$\liminf_{k\to\infty} A^M_{n_k}(\pi_k)0 \geqslant A^M(\pi)0$$

which was needed here above.

Since $\varepsilon > 0$ was arbitrary we obtain (*a*) of Definition 3.3. This completes the proof.

The preceding proof also implies a further result on the mode of convergence.

3.6. Theorem

Under the assumptions in 2.1 and 3.1 the value functions v_n, $n \geqslant 1$ converge pointwise to v and the convergence is monotone non-decreasing.

We note the Theorem 3.6 suffices for v_n, $n \geqslant 1$ to epi-converge to v. We can now sum up the proof of Theorem, 3.2.

As mentioned, statement (i) follows directly from Theorems 3.4 and 3.5. Statement (ii) is a result of the compactness of Π.

References

Attouch, H., and Wets, R. J.-B., (1983), Approximation and convergence in nonlinear optimization. *Nonlinear Programming*, **4**, eds. 0. Mangasarian *et al* (Academic Press, New York).

Attouch, H., and Wets, R. J.-B., (1983), A convergence theory for saddle functions. *Trans. Am. Math. Soc.*, **280**, 1–49.

Bertsekas, D., (1976), *Dynamic Programming and Stochastic Control.* (Academic Press, New York).

Bertsekas, D., and Shreve, S., (1978), *Stochastic Optimal Control: The Discrete Time Case.* (Academic Press, New York).

Denardo, E., (1982), *Dynamic Programming.* (Prentice-Hall, New York).

Flåm, S. D, (1987), Approximating Some Convex Programs in terms of Borel Fields, to appear in Mathematical Programming Study (31).

Fox, B., (1971), Finite-state approximations to denumerable-state dynamic programs. *J. Math. Anal. App.*, **34**, 675–670.

Gihman, I. I., and Skorohod, A. V., (1979), *Controlled Stochastic Processes.* (Springer-Verlag, New York).

Kall, P., (1986), Approximation to optimization problems: an elementary review. *Mathematics of Operations Research*, **11**, 9–18.

Kall, P., (1987), Stochastic Programs with Recourse: An Upper Bound and the Related Moment Problem. Tech. Report, Institut für Operations Research der Universität Zürich.

MacQueen, J., (1967), A test for suboptimal actions in Markovian decision problems. *Operations Research*, **15**, 559–561.

MENDELSSOHN, R., (1980), Improved bounds for aggregated linear programs. *Operations Research*, **28**, 1450–1453.

NORMAN, J., and WHITE, D., (1968), A method for approximating solutions to stochastic dynamic programming problems using expectations. *Operations Research*, **16**, 296–306.

ROSS, S., (1983), *Introduction to Stochastic Dynamic Programming*. (Academic Press, New York).

WHITE, D., (1980), Finite-state approximations for denumerable-state infinite-horizon discounted Markov decision processes. *J. Math. Anal. App.*, **74**, 292–295.

WHITT, W., (1978, 1979), Approximation of dynamic programs, I and II, *Mathematics of Operations Research*, **3**, 231–243, 175–185.