stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE BW 20/73 APRIL

A. HORDIJK

ON THE CONVERGENCE OF THE AVERAGE EXPECTED RETURN IN DYNAMIC PROGRAMMING

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS (MOS) subject classification scheme (1970): 90C40

On the convergence of the average expected return in dynamic programming

by

Arie Hordijk

Abstract

Under a certain condition it is shown that the average expected return in dynamic programming converges.
The proof uses a sequence of contraction mappings.

ON THE CONVERGENCE OF THE AVERAGE EXPECTED RETURN IN DYNAMIC PROGRAMMING.

Arie Hordijk

Suppose we have a dynamic programming problem with state space S, action or decision space A, law of motion q and bounded return function r. Under general conditions the optimal α-discounted return v_α satisfies the functional equation (see [1])

$$(1) \qquad v_\alpha(x) = \sup_{a\in A} \{r(x,a) + \alpha\int_S q(dy|x,a)v_\alpha(y)\}.$$

Define $w_0(x) \equiv 0$ and

$$(2) \qquad w_{n+1}(x) = \sup_{a\in A} \{r(x,a) + \int_S q(dy|x,a)w_n(y)\}$$

The sequence w_n is a dynamic programming sequence.
w_n represents the optimal return in n periods. It is well-known that in the finite state and action model w_n/n converges to the optimal average return (see [3]).
We assume the existence of constants c and α_0 such that

$$(3) \qquad |(1-\alpha_1)v_{\alpha_1}(x) - (1-\alpha_2)v_{\alpha_2}(x)| \leq |\alpha_1-\alpha_2|c, \text{ for all } \alpha_0 < \alpha_1, \alpha_2 < 1$$

and all $x\in S$.

This means that v_α has a partial Laurent series expansion and consequently $\lim_{\alpha\to 1} (1-\alpha)v_\alpha$ exists and is finite. Using a sequence of contraction mappings, we shall prove that assumption (3) implies $\lim_{n\to\infty} w_n/n = \lim_{\alpha\to 1} (1-\alpha)v_\alpha$.

Proof. Let $\alpha_n = 1 - 1/n$ then for k_0 such that $\alpha_{k_0} > \alpha_0$

(4) $$\prod_{k=k_0+1}^{n} \to 0 \quad \text{and} \quad \sum_{k=k_0+1}^{n} \prod_{j=k}^{n} \alpha_j(\alpha_k-\alpha_{k-1}) \to 0 \quad \text{as } n \to \infty$$

Define the contraction mapping T_n by

(5) $$(T_n g)(x) = \sup_{a\in A} \{r(x,a)/n + (1-1/n)\int_S q(dy|x,a)g(y)\}$$

It then follows from (1) that $(1-\alpha_n)v_{\alpha_n}$ is a fixed-point of T_n i.e.

(6) $$T_n[(1-\alpha_n)v_{\alpha_n}] = (1-\alpha_n)v_{\alpha_n}$$

Relation (2) implies

(7) $$T_n[w_{n-1}/n-1] = w_n/n$$

From (6) and (7) and the fact that T_n has contraction-modulus α_n it follows that

(8) $$\| w_n/n -(1-\alpha_n)v_{\alpha_n} \| \le \alpha_n \| w_{n-1}/n-1 -(1-\alpha_n)v_{\alpha_n} \| ,$$

where $\|g\|$ denotes $\sup_{x\in S} |g(x)|$.

By using the triangle inequality we deduce from (3) and (8)

(9) $$\| w_n/n -(1-\alpha_n)v_{\alpha_n} \| \le \alpha_n \| w_{n-1}/n-1 -(1-\alpha_{n-1})v_{\alpha_{n-1}} \| + \alpha_n(\alpha_n-\alpha_{n-1})c$$

Iterating this inequality, we find

$$(10)\qquad \| w_n/n - (1-\alpha_n)v_n \| \le \prod_{k=k_0+1}^{n} \alpha_k \| w_{k_0}/k_0 - (1-\alpha_{k_0})v_{k_0} \| + \sum_{k=k_0+1}^{n} \prod_{j=k}^{n} \alpha_j (\alpha_k - \alpha_{k-1}) c$$

From (4) it follows then

$$\lim_{n\to\infty} \| w_n/n - (1-\alpha_n)v_{\alpha_n} \| = 0$$

and consequently

$$\lim_{n\to\infty} w_n/n = \lim_{n\to\infty} (1-\alpha_n)v_{\alpha_n} \ . \ \Box$$

To conclude we show that in the finite state and action model the function $(1-\alpha)v_\alpha$ has a bounded derivative for α sufficiently near 1 from which it follows that assumption (3) is satisfied.

In the finite case there exists a Blackwell-optimal policy i.e. a stationary policy which is discounted-optimal for all discount-factors $\alpha_0 < \alpha < 1$ for some α_0 (see [2]). Using the Laurent series expansion as given by Miller and Veinott (see theorem 1 of [4]) we find

$$(11)\qquad (1-\alpha)v_\alpha = \sum_{n=0}^{\infty} \rho^n y_n \text{ , with } \rho = \alpha^{-1}(1-\alpha), \quad y_0 = P^*(f)r(f) \text{ and}$$

$$y_n = (-1)^{n-1} H(f)^n r(f), \quad n=1,2,\ldots, \text{ for } f \text{ a Blackwell-optimal poli-}$$

cy.

Since the series in (11) converges for all $(\rho) < \|H(f)\|^{-1}$, it follows that $(1-\alpha)v_\alpha$ has a bounded derivative with respect to ρ and consequently also the derivative with respect to α is bounded for α sufficiently near 1.

REFERENCES

1. R. BELLMAN, *Dynamic Programming*,
Princeton University Press, Princeton, 1957.

2. D. BLACKWELL, Discrete dynamic programming,
Ann. Math. Statist. 33 (1962) 719-726.

3. C. DERMAN, *Finite State Markovian Decision Processes*,
Academic Press, New York, 1970.

4. B.L. MILLER and A.F. VEINOTT, Discrete dynamic programming with a small interest rate,
Ann. Math. Statist. 40 (1969) 336-370.