Reprinted from JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS Vol. 46, No. 2, May 1974
All Rights Reserved by Academic Press, New York and London *Printed in Belgium*

On the Convergence of the Average Expected Return in Dynamic Programming*

ARIE HORDIJK

Foundation Mathematisch Centrum, 49 Amsterdam, Holland

Submitted by Richard Bellman

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])

$$v_\alpha(x) = \sup_{a \in A} \left\{ r(x, a) + \alpha \int_S q(dy \mid x, a)\, v_\alpha(y) \right\}. \tag{1}$$

Define $w_0(x) \equiv 0$ and

$$w_{n+1}(x) = \sup_{a \in A} \left\{ r(x, a) + \int_S q(dy \mid x, a)\, w_n(y) \right\}. \tag{2}$$

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

$$\text{for all} \quad \begin{gathered} |(1 - \alpha_1)\, v_{\alpha_1}(x) - (1 - \alpha_2)\, v_{\alpha_2}(x)| \leqslant |\, \alpha_1 - \alpha_2 \,|\, c, \\ \alpha_0 < \alpha_1\,, \alpha_2 < 1, \quad \text{and all} \quad x \in S. \end{gathered} \tag{3}$$

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 $\alpha_{k_0} > \alpha_0$

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

* Report BW 20/73 of the Mathematical Centre, Amsterdam.

Define the contraction mapping T_n by

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

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

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

Relation (2) implies

$$T_n[w_{n-1}/n - 1] = w_n/n. \tag{7}$$

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

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

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

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

$$\begin{aligned} &\| w_n/n - (1 - \alpha_n)\, v_{\alpha_n} \| \\ &\quad \leqslant \alpha_n \| w_{n-1}(n - 1)^{-1} - (1 - \alpha_{n-1})\, v_{\alpha_{n-1}} \| + \alpha_n(\alpha_n - \alpha_{n-1})\, c. \end{aligned} \tag{9}$$

Iterating this inequality, we find

$$\begin{aligned} &\| w_n/n - (1 - \alpha_n)\, v_n \| \\ &\quad \leqslant \prod_{k=k_0+1}^{n} \alpha_k \| w_{k_0}/k_0 - (1 - \alpha_{k_0})\, v_{k_0} \| + \sum_{k=k_0+1}^{n} \prod_{j=k}^{n} \alpha_j(\alpha_k - \alpha_{k-1})\, c. \end{aligned} \tag{10}$$

From (4) it follows then that

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

and consequently

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

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

$$(1 - \alpha)\, v_\alpha = \sum_{n=0}^{\infty} \rho^n y_n\,, \tag{11}$$

with $\rho = \alpha^{-1}(1 - \alpha)$, $y_0 = P^*(f)\, r(f)$, and

$$y_n = (-1)^{n-1}\, H(f)^n\, r(f) \qquad (n = 1, 2,...)$$

for f a Blackwell optimal policy. 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, NJ, 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, Jr., Discrete dynamic programming with a small interest rate, *Ann. Math. Statist.* **40** (1969), 336–370.

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.