

STICHTING  
MATHEMATISCH CENTRUM  
2e BOERHAAVESTRAAT 49  
AMSTERDAM

MAITRA A

A Note on Undiscounted Dynamic Programming  
S 355 (SP89)

The Annals of Mathematical Statistics,  
37(1966) II, p 1042-1044.



1965

# A NOTE ON UNDISCOUNTED DYNAMIC PROGRAMMING<sup>1</sup>

BY ASHOK MAITRA<sup>2</sup>

*Mathematisch Centrum, Amsterdam*

1. Introduction. We consider a system with a finite number of states  $1, 2, \dots, S$ . Once a day, we observe the current state  $s$  of the system and choose an action  $a$  from an arbitrary set  $A$  of actions. As a result, two things happen: (1) we receive an immediate income  $i(s, a)$ , and (2) the system moves to a new state  $s'$  with probability  $q(s' | s, a)$ . Assume that the incomes are bounded, that is, there exists a positive number  $M$  such that  $|i(s, a)| \leq M$ ,  $s = 1, 2, \dots, S$ ,  $a \in A$ . The problem is to maximise the average rate of income (to be defined below).

Denote by  $F$  the set of all functions  $f$  on  $S$  into  $A$ . A *policy*  $\pi = \{f_1, f_2, \dots\}$  is a sequence of functions  $f_n \in F$ . Thus, to use policy  $\pi$  is to choose the action  $f_n(s)$  on the  $n$ th day, if the system is in state  $s$  on that day. We shall call a policy  $\pi = \{f_n\}$  *stationary* if  $f_n = f$ ,  $n = 1, 2, \dots$ , and denote it by  $f^{(\infty)}$ .

With each  $f \in F$ , associate (1) the  $S \times 1$  vector  $r(f)$ , whose  $s$ th coordinate is  $i(s, f(s))$  and (2) the  $S \times S$  stochastic matrix  $Q(f)$ , whose  $(s, s')$  element is  $q(s' | s, f(s))$ . Hence, if we use the policy  $\pi = \{f_n\}$ , the  $n$ -step transition matrix of the system is  $Q_n(\pi) = \prod_{k=1}^n Q(f_k)$ . In particular, if our policy is stationary, the system becomes a discrete time-parameter Markov chain with stationary transition probabilities.

Given a policy  $\pi$ , let us denote by  $W_n(\pi)$  the  $S \times 1$  vector of incomes on the  $n$ th day, when the policy  $\pi$  is used. Set

$$x(\pi) = \lim_{N \rightarrow \infty} N^{-1} \sum_{n=1}^N W_n(\pi)$$

whenever the limit exists. Blackwell [1] has shown that the limit exists whenever  $\pi$  is a stationary policy. In the case of a stationary policy,  $x(f^{(\infty)})$  is the vector of average rates of income, when the policy  $f^{(\infty)}$  is used.

We shall say that a policy  $f_0^{(\infty)}$  is *optimal* among stationary policies if  $x(f_0^{(\infty)}) \geq x(f^{(\infty)})$  for all  $f \in F$  (for any two  $S \times 1$  vectors  $w_1$  and  $w_2$ , we shall write  $w_1 \geq w_2$  if every coordinate of  $w_1$  is at least as large as the corresponding coordinate of  $w_2$ , and  $w_1 > w_2$  if  $w_1 \geq w_2$  and  $w_1 \neq w_2$ ).

Blackwell [1] showed that, if  $A$  is finite, there exists an optimal policy among stationary policies. When  $A$  is not finite, there may not exist an optimal policy. Consider, for instance, a system with a single state and  $A = \{1, 2, \dots\}$ . Choice of action  $i$  brings an income of  $1 - 1/i$  dollars. It is clear that there is no optimal stationary policy.

The purpose of this note is to prove:

**THEOREM.** *Let  $A$  be arbitrary. Given  $\epsilon > 0$ , there exists a stationary policy  $f_\epsilon^{(\infty)}$*

Received 24 November 1965.

<sup>1</sup> Report SP 89 of the Statistics Department, Mathematisch Centrum, Amsterdam.

<sup>2</sup> Now with Indian Statistical Institute, Calcutta.

such that  $x(f_\epsilon^{(\infty)}) \geq \sup_{f \in F} x(f^{(\infty)}) - \epsilon e$ , where  $e$  is the  $S \times 1$  vector with all coordinates unity.

2. **Proof of theorem.** We introduce a discount factor  $\beta$ ,  $0 \leq \beta < 1$ , so that the value of unit income  $n$  days in the future is  $\beta^n$ . Blackwell [1] has shown that the total expected discounted return from a policy  $f^{(\infty)}$  is given by the  $S \times 1$  vector

$$V_\beta(f^{(\infty)}) = \sum_{n=0}^{\infty} \beta^n [Q(f)]^n r(f)$$

and that

$$x(f^{(\infty)}) = \lim_{\beta \rightarrow 1} (1 - \beta) V_\beta(f^{(\infty)}).$$

With each  $f \in F$  and each  $\beta$ ,  $0 \leq \beta < 1$ , let us associate the transformation  $L_\beta(f)$  which maps the  $S \times 1$  vector  $w$  into  $L_\beta(f)w = r(f) + \beta Q(f)w$ . We note that  $L_\beta(f)$  is monotone, that is,  $w_1 \geq w_2$  implies  $L_\beta(f)w_1 \geq L_\beta(f)w_2$ . Note that  $V_\beta(f^{(\infty)})$  is the fixed point of  $L_\beta(f)$ .

In order to prove our theorem, we need a lemma.

LEMMA. Let  $f_1, f_2, \dots, f_k \in F$  ( $k \geq 2$ ). Then there exists  $h \in F$  such that

$$V_\beta(h^{(\infty)}) \geq V_\beta(f_i^{(\infty)}), \quad i = 1, 2, \dots, k$$

for all  $\beta \geq$  some  $\beta_0$ .

PROOF. It suffices to prove the lemma for  $k = 2$ . The proof for general  $k$  then proceeds by induction.

Denote by  $u_s$  the  $s$ th coordinate of the  $S \times 1$  vector  $u$ .

Consider  $V_\beta(f_1^{(\infty)})_s$  and  $V_\beta(f_2^{(\infty)})_s$ . Either  $V_\beta(f_1^{(\infty)})_s \geq V_\beta(f_2^{(\infty)})_s$  for all  $\beta \geq$  some  $\beta'$  or  $V_\beta(f_1^{(\infty)})_s < V_\beta(f_2^{(\infty)})_s$  for a sequence of  $\beta$ 's tending to 1. But for each  $s$  and each  $f$ ,  $V_\beta(f^{(\infty)})_s$  is a rational function of  $\beta$ , as the representation  $V_\beta(f^{(\infty)}) = [I - \beta Q(f)]^{-1} r(f)$  shows. Consequently, either  $V_\beta(f_1^{(\infty)})_s \geq V_\beta(f_2^{(\infty)})_s$  for all  $\beta \geq$  some  $\beta''$  or  $V_\beta(f_1^{(\infty)})_s < V_\beta(f_2^{(\infty)})_s$  for all  $\beta \geq$  some  $\beta''$ . Thus, for each  $s$ , there exists a  $\beta_s < 1$  such that either  $V_\beta(f_1^{(\infty)})_s \geq V_\beta(f_2^{(\infty)})_s$  for all  $\beta \geq \beta_s$  or  $V_\beta(f_1^{(\infty)})_s < V_\beta(f_2^{(\infty)})_s$  for all  $\beta \geq \beta_s$ .

Let  $\beta_0 = \max_{1 \leq s \leq S} \beta_s$ . For each  $\beta \geq \beta_0$ , define  $u(\beta)_s = \max(V_\beta(f_1^{(\infty)})_s, V_\beta(f_2^{(\infty)})_s)$ . We now define  $h \in F$  as follows:

$$\begin{aligned} h(s) &= f_1(s) && \text{if } V_\beta(f_1^{(\infty)})_s \geq V_\beta(f_2^{(\infty)})_s \text{ for all } \beta \geq \beta_0 \\ &= f_2(s) && \text{if } V_\beta(f_1^{(\infty)})_s < V_\beta(f_2^{(\infty)})_s \text{ for all } \beta \geq \beta_0, \quad 1 \leq s \leq S. \end{aligned}$$

Set  $u(\beta) = (u(\beta)_1, u(\beta)_2, \dots, u(\beta)_S)$ . It is easy to check that  $L_\beta(h)u(\beta) \geq u(\beta)$  for all  $\beta \geq \beta_0$ . Denoting by  $L_\beta^{(n)}(h)$  the  $n$ th iterate of  $L_\beta(h)$ , we see that  $L_\beta^{(N)}(h)u(\beta) \geq u(\beta)$  for  $N = 1, 2, \dots$  and all  $\beta \geq \beta_0$ . For fixed  $\beta \geq \beta_0$ , let  $N \rightarrow \infty$ . We get:  $V_\beta(h^{(\infty)}) \geq u(\beta)$  for all  $\beta \geq \beta_0$ . This completes the proof of the lemma.

PROOF OF THEOREM. Set  $x_s^* = \sup_{f \in F} (x(f^{(\infty)}))_s$  and  $x^* = (x_1^*, x_2^*, \dots, x_S^*)$ . Let  $\epsilon > 0$ . For each  $s$ , choose  $f_s \in F$  such that  $x(f_s^{(\infty)})_s > x_s^* - \epsilon$ . Hence, for each  $s$ , there exists  $\beta_s' < 1$  such that  $(1 - \beta)V_\beta(f_s^{(\infty)})_s > x_s^* - \epsilon$  for all  $\beta \geq$

$\beta_s'$ . Let  $\beta' = \max_{1 \leq s \leq S} \beta_s'$ . But by the preceding lemma, there exists  $h \in F$  and  $\beta'' < 1$  such that  $V_\beta(h^{(\infty)}) \geq V_\beta(f_s^{(\infty)})$  for  $1 \leq s \leq S$  and all  $\beta \geq \beta''$ . Hence  $(1 - \beta)V_\beta(h^{(\infty)}) > x^* - \epsilon$  for all  $\beta \geq \max(\beta', \beta'')$ . Let  $\beta \rightarrow 1$ . We get:  $x(h^{(\infty)}) \geq x^* - \epsilon$ . The proof is completed by taking  $h = f_\epsilon$ .

REMARK. In [2], I gave an example of a system with countably infinite state space and finite action space  $A$ , where there exists no optimal policy among stationary policies. It would be of interest to know if there exist  $\epsilon$ -optimal policies in this case.

## REFERENCES

- [1] BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719-726.
- [2] MAITRA, A. (1965). Dynamic programming for countable state systems. *Sankhyā Ser. A* **27** 259-266.