A NOTE ON UNDISCOUNTED DYNAMIC PROGRAMMING[1]

By Ashok Maitra[2]

Mathematisch Centrum, Amsterdam

1. Introduction. We consider a system with a finite number of states $1, 2, \cdots, S$. 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 $q(s' \mid s, a)$. Assume that the incomes are bounded, that is, there exists a positive number M such that $|i(s, a)| \leqq M$, $s = 1, 2, \cdots, S$, $a \varepsilon A$. 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* $\pi = \{f_1, f_2, \cdots\}$ is a sequence of functions $f_n \varepsilon F$. Thus, to use policy π is to choose the action $f_n(s)$ on the nth day, if the system is in state s on that day. We shall call a policy $\pi = \{f_n\}$ *stationary* if $f_n = f$, $n = 1, 2, \cdots$, and denote it by $f^{(\infty)}$.

With each $f \varepsilon F$, associate (1) the $S \times 1$ vector $r(f)$, whose sth coordinate is $i(s, f(s))$ and (2) the $S \times S$ stochastic matrix $Q(f)$, whose (s, s') element is $q(s' \mid s, f(s))$. 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 π, let us denote by $W_n(\pi)$ the $S \times 1$ vector of incomes on the nth day, when the policy π is used. Set

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

whenever the limit exists. Blackwell [1] has shown that the limit exists whenever π 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 $x(f_0^{(\infty)}) \geqq x(f^{(\infty)})$ for all $f \varepsilon F$ (for any two $S \times 1$ vectors w_1 and w_2, we shall write $w_1 \geqq w_2$ if every coordinate of w_1 is at least as large as the corresponding coordinate of w_2, and $w_1 > w_2$ if $w_1 \geqq w_2$ 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 $A = \{1, 2, \cdots\}$. 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.

[1] Report SP 89 of the Statistics Department, Mathematisch Centrum, Amsterdam.

[2] Now with Indian Statistical Institute, Calcutta.

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

2. Proof of theorem. We introduce a discount factor β, $0 \leqq \beta < 1$, so that the value of unit income n days in the future is β^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

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

With each $f \varepsilon F$ and each β, $0 \leqq \beta < 1$, 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, $w_1 \geqq w_2$ implies $L_\beta(f)w_1 \geqq L_\beta(f)w_2$. 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* $f_1, f_2, \cdots, f_k \varepsilon F$ $(k \geqq 2)$. *Then there exists* $h \varepsilon F$ *such that*

$$V_\beta(h^{(\infty)}) \geqq V_\beta(f_i^{(\infty)}), \qquad i = 1, 2, \cdots, k$$

for all $\beta \geqq$ *some* β_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 sth coordinate of the $S \times 1$ vector u.

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

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

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

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

Proof of Theorem. Set $x_s^* = \sup_{f \varepsilon F} (x(f^{(\infty)})_s)$ and $x^* = (x_1^*, x_2^*, \cdots, x_S^*)$. Let $\epsilon > 0$. For each s, choose $f_s \varepsilon F$ 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 $\beta \geqq$

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

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 ϵ-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.