POSITIVE DYNAMIC PROGRAMMING

DAVID BLACKWELL
UNIVERSITY OF CALIFORNIA, BERKELEY

1. Introduction

A dynamic programming problem is specified by four objects: S, A, q, r, where S is a nonempty Borel set, the set of *states* of some system, A is a nonempty Borel set, the set of *acts* available to you, q is the *law of motion* of the system; it associates (Borel measurably) with each pair (s, a) a probability distribution $q(\cdot \mid s, a)$ on S: when the system is in state s and you choose act a, the system moves to a new state selected according to $q(\cdot \mid s, a)$, and r is a bounded Borel measurable function on $S \times A \times S$, the *immediate return:* when the system is in state s, and you choose act a, and the system moves to s', you receive an income $r(s, a, s')$. A *plan* π is a sequence $\pi_1, \pi_2, \cdots$, where π_n tells you how to select an act on the n-th day, as a function of the previous history $h = (s_1, a_1, \cdots, a_{n-1}, s_n)$ of the system, by associating with each h (Borel measurably) a probability distribution $\pi_n(\cdot \mid h)$ on (the Borel subsets of) A.

Any sequence of Borel measurable functions $f_1, f_2, \cdots$, each mapping S into A, defines a plan. When in state s on the n-th day, choose act $f_n(s)$. Plans $\pi = \{f_n\}$ of this type may be called *Markov* plans. A single f defines a still more special kind of plan: whenever in state s, choose act $f(s)$. This plan is denoted by $f^{(\infty)}$, and plans $f^{(\infty)}$ are called *stationary*.

A plan π associates with each initial state s a corresponding *expected* n-th *period return* $r_n(\pi)(s)$ and an *expected discounted total return*

$$I_\beta(\pi)(s) = \sum_1^\infty \beta^{n-1} r_n(\pi)(s), \tag{1}$$

where β is a fixed discount factor, $0 \leq \beta < 1$.

The problem of finding a π which maximizes I_β was studied in [1]. Three of the principal results obtained were the following.

RESULT (i). *For any probability distribution p on S and any $\epsilon > 0$, there is a stationary plan $f^{(\infty)}$ which is (p, ϵ)-optimal; that is,*

$$p\{I_\beta(f^{(\infty)}) > I_\beta(\pi) - \epsilon\} = 1 \qquad \text{for all } \pi. \tag{2}$$

RESULT (ii). *Any bounded u which satisfies*

$$u(s) \geq \int [r(s, a, \cdot) + \beta u(\cdot)]\, dq(\cdot \mid s, a) \qquad \text{for all } s, a \tag{3}$$

is an upper bound on incomes;

This research was supported by the Information Systems Branch of the Office of Naval Research under Contract Nonr-222(53).

$$I_\beta(\pi) \leq u \qquad \text{for all } \pi. \tag{4}$$

RESULT (iii). *If A is countable, the optimal return u_β^* is the unique bounded fixed point of the operator U_β, mapping the set of bounded functions u on S into itself, defined by*

$$U_\beta u(s) = \sup_a \int [r(s, a, \cdot) + \beta u(\cdot)]\, dq(\cdot \mid s, a), \tag{5}$$

that is, $u_\beta^* = \sup_\pi I_\beta(\pi)$. Also $U_\beta^n u \to u_\beta^*$ as $n \to \infty$, for every bounded u.

In this paper we consider the positive (undiscounted) case $r \geq 0$, $\beta = 1$, and we are interested in maximizing $I(\pi) = \sum r_n(\pi)$. A weakened form of (i) is proved. Modified forms of (ii) and (iii) are obtained.

THEOREM 1. *For any probability distribution p on S for which*

$$v = \sup_\pi \int I(\pi)\, dp \tag{6}$$

is finite, and any $\epsilon > 0$, there is a stationary plan $f^{(\infty)}$ which is weakly (p, ϵ)-optimal, that is

$$\int I(f^{(\infty)})\, dp > v - \epsilon. \tag{7}$$

THEOREM 2 (Compare [2], theorem 2.12.1). *Any nonnegative u which satisfies*

$$u(s) \geq \int [r(s, a, \cdot) + u(\cdot)]\, dq(\cdot \mid s, a) \qquad \text{for all } s, a \tag{8}$$

is an upper bound on incomes;

$$I(\pi) \leq u \qquad \text{for all } \pi. \tag{9}$$

THEOREM 3. *If A is countable, the optimal return u^* is the smallest nonnegative fixed point of the operator U, taking the set of nonnegative (possibly $+\infty$-valued) functions on S into itself, defined by*

$$Uu(s) = \sup_a \int [r(s, a, \cdot) + u(\cdot)]\, dq(\cdot \mid s, a). \tag{10}$$

Also $U^n 0 \to u^*$ as $n \to \infty$.

2. Proofs

Theorem 1 is an easy consequence of (i): first choose π so that $\int I(\pi)\, dp > v - \epsilon$; next, choose $\beta < 1$ so that $\int I_\beta(\pi)\, dp > v - \epsilon$. Now invoke (i) to choose $f^{(\infty)}$ in such a way that

$$p\{I_\beta(f^{(\infty)}) > I_\beta(\pi) - \delta\} = 1, \tag{11}$$

where $\delta = \int I_\beta(\pi)\, dp + \epsilon - v$, so that

$$\int I(f^{(\infty)})\, dp \geq \int I_\beta(f^{(\infty)})\, dp > \int I_\beta(\pi)\, dp - \delta = v - \epsilon. \tag{12}$$

Similarly, theorem 2 is an easy consequence of (ii): fix $\beta < 1$, and define $w = \min(u, R/(1 - \beta))$, where $R = \sup_{s,a,s'} r(s, a, s')$. We show that w satisfies the hypothesis of (ii), that is,

$$w(s) \geq \int [r(s, a, \cdot) + \beta w(\cdot)]\, dq(\cdot \mid s, a) \qquad \text{for all } s, a. \tag{13}$$

First,

$$\begin{aligned} u(s) &\geq \int [r(s, a, \cdot) + u(\cdot)]\, dq(\cdot \mid s, a) \\ &\geq \int [r(s, a, \cdot) + w(\cdot)]\, dq(\cdot \mid s, a). \end{aligned} \tag{14}$$

Second, putting $R/(1 - \beta) = c$,

$$c = R + \beta c \geq \int [r(s, a, \cdot) + \beta w(\cdot)]\, dq(\cdot \mid s, a). \tag{15}$$

Thus,

$$w = \min(u, c) \geq \int [r(s, a, \cdot) + \beta w(\cdot)]\, dq(\cdot \mid s, a). \tag{16}$$

So, (ii) implies that $I_\beta(\pi) \leq u$ for all π, $\beta < 1$. Letting $\beta \to 1$ yield $I(\pi) \leq u$.

For theorem 3, note that U is monotone: if $u \geq v$, then $Uu \geq Uv$. Hence, $U^n 0 = u_n$ increases with n, say to w. We show that w is a fixed point of U. Define the operator T_a by

$$T_a u(s) = \int [r(s, a, \cdot) + u(\cdot)]\, dq(\cdot \mid s, a), \tag{17}$$

so that $U = \sup_a T_a$. We have $T_a u_n \leq U u_n = u_{n+1} \leq Uw$, so that $(n \to \infty)$ $T_a w \leq w \leq Uw$ and $(\sup_a)$ $Uw \leq w \leq Uw$. The function w is the smallest non-negative fixed point of U, since $v \geq 0$ and $Uv = v$ imply, applying U n times to $v \geq 0$, $v \geq U^n 0$, so $(n \to \infty)$ $v \geq w$.

To identify the optimal return with w, note that, for any $\beta < 1$, we have $U_\beta^n 0 \leq U^n 0$, so that $(n \to \infty)$ $u_\beta \leq w$. Thus,

$$I_\beta(\pi) \leq w \quad \text{for all } \pi, \quad \text{and} \quad I(\pi) \leq w \quad \text{for all } \pi. \tag{18}$$

On the other hand,

$$\begin{aligned} &\sup_\pi I(\pi) \geq \sup_\pi I_\beta(\pi) = u_\beta^* \geq U_\beta^n(0), \quad \text{so that} \quad (\beta \to 1), \\ &\sup_\pi I(\pi) \geq U^n 0; \quad \text{and} \quad (n \to \infty), \sup_\pi I(\pi) \geq w. \end{aligned} \tag{19}$$

3. Remarks

(1) In the negative case, $r \leq 0$, $\beta = 1$, Dubins and Savage [2] have given an example in which theorem 1 is false. Ralph Strauch has recently studied the negative case extensively in his thesis, finding that it differs substantially in other ways from the positive case.

(2) Here is an example in which no (p, ϵ)-optimal stationary plan exists, showing that (i) cannot be generally extended to the positive case. There is a sequence $p(1), p(2), \cdots$ of primary states, a sequence $s(1), s(2), \cdots$ of secondary states, and a terminal state t. From primary state $p(n)$ we have two choices: (1) move to secondary state $s(2^n - 1)$, or (2) move to the next primary state $p(n + 1)$ with probability $\frac{1}{2}$, and to the terminal state t with probability $\frac{1}{2}$.

The immediate income is 0 no matter what happens. From secondary state $s(n)$, $n \geq 2$, you move to secondary state $s(n-1)$ and receive \$1. In secondary state $s(1)$ you move to t and receive \$1. Once state t is reached, you stay there and receive nothing.

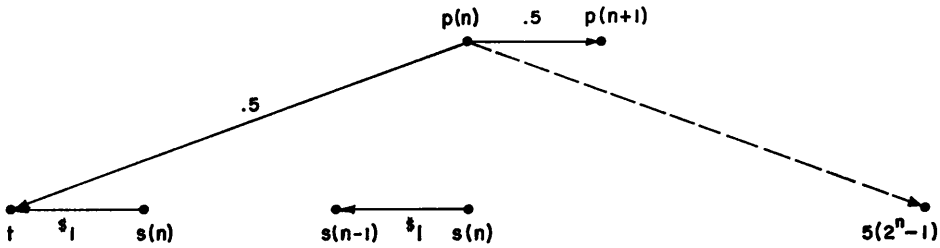

FIGURE 1

The income from $s(n)$ is n, and from t is 0. From $p(n)$, by aiming for $s(2^{n+k}-1)$ via $p(n+k)$, your expected income is $(2^{n+k}-1)/2^k = 2^n - 2^{-k}$, so you can get nearly 2^n from $p(n)$. The function u: $u(p(n)) = 2^n$, $u(s(n)) = n$, $u(t) = 0$ satisfies the hypothesis of theorem 2 (with equality), so is the optimal return: $u = \sup_\pi I(\pi)$. But for any stationary plan $f^{(\infty)}$, either f elects to gamble at every primary state so that $I(f^{(\infty)}) = 0$ for all $p(n)$, or there is a primary state $p(n_0)$ from which f moves to $s(2^{n_0}-1)$ so that $I(f^{(\infty)}) = 2^{n_0} - 1$ at $p(n_0)$, one dollar short of the optimal return at $p(n_0)$. So for any p which assigns positive probability to every primary state and any $\epsilon < 1$, there is no (p, ϵ)-optimal stationary plan.

(3) In the above example, the optimal return is unbounded. Don Ornstein (unpublished) has shown that for a certain class of (positive) problems with bounded optimal return and countable state space, there is for every $\epsilon > 0$ an ϵ-optimal plan $f^{(\infty)}$ which is stationary:

$$I(f^{(\infty)}) > I(\pi) - \epsilon \qquad \text{for all } \pi, s. \tag{20}$$

His method appears to apply to any (positive) problem with bounded optimal return and countable state space. Whether there is a (p, ϵ)-optimal stationary plan in every positive problem with bounded optimal return remains open.

(4) In the discounted case, if there is an optimal plan, there is one which is stationary. Whether this is true in the positive case remains open, even for bounded optimal return.

REFERENCES

[1] DAVID BLACKWELL, "Discounted dynamic programming," *Ann. Math. Statist.*, Vol. 36 (1965), pp. 226–235.

[2] L. E. DUBINS and L. J. SAVAGE, *How to Gamble If You Must*, New York, McGraw-Hill, 1965.