JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS **116**, 594–597 (1986)

Fixed Point Theorems for Discounted Finite Markov Decision Processes

ULRICH DIETER HOLZBAUR

Abt. Math. VII (Operations Research), Universität Ulm, D-7900 Ulm, West Germany

Submitted by E. Stanley Lee

We establish the existence of a solution to the optimality equation for discounted finite Markov decision processes by means of Birkhoff's fixed point theorem. The proof yields the well-known linear programming formulation for the optimal value function while its dual characterizes the optimal value function as the maximum over all value functions.

For a discounted finite state and action Markov decision process (e.g., see [1, 2]), we call any solution V of the optimality equation

$$V_i = \max_{a=1\ldots A} \left(r(i,a) + \beta \sum_{j=1}^{S} p_{ij}(a)\, V_j \right), \qquad i = 1 \cdots S \tag{OE}$$

an optimal value function. Introducing the optimal reward operator

$$T\colon \quad \mathbb{R}^S \to \mathbb{R}^S$$

$$v \to \left(\max_{a=1\ldots A} \left(r(i,a) + \beta \sum_{j=1}^{S} p_{ij}(a)\, v_j \right) \right), \qquad i = 1 \cdots S$$

we may write (OE) as $V = TV$.

The usual way to prove the existence of an optimal value function is to apply Banach's fixed point theorem to the contraction mapping T (e.g., see [3, 4]). The method of successive approximations applied to T is the value iteration of dynamic programming, which is not finite but returns the optimal policy of the decision process in a finite number of steps (turnpike theorems, cf. [5, 6]).

Shapiro showed in [7] that V can also be obtained by use of Brouwer's fixed point theorem. We do not know whether the algorithm of Scarf [8], proposed in [7] for the computation of the fixed point of T, has any advantage over the value iteration.

We give the proof for the existence of a solution to the optimality

0022-247X/86 $3.00

equation that uses the monotonicity of T instead of its continuity. Moreover, we show that this access is equivalent to the linear programming approach to Markov decision processes used first by D'Epenoux [9] and Manne [10] and studied extensively by Kallenberg [11].

First we give a formulation and a proof of Birkhoff's fixed point theorem [12, 13]:

Let L be a complete lattice and $T: L \to L$ isotone. Then there exists $V \in L$ such that $TV = V$.

Proof. Let $S = \{s \in L: s \geqq Ts\}$, $V = \inf S$. Then for any $s \in S$ we have $s \geqq Ts \geqq TV$, which implies that $V = \inf S \geqq TV$ and $V \in S$. Now $TV \geqq T(TV)$ implies $TV \in S$, hence $TV \geqq V$.

The existence of an optimal value function now follows from Birkhoff's theorem and the following three lemmata. For this, let $M = \max_{i,a} |r(i, a)|$, $\lambda = M/(1-\beta)$, $l = (\lambda, \lambda, ..., \lambda) \in \mathbb{R}^S$, $L = \{x \in \mathbb{R}^S: |x_i| \leqq \lambda,\ i = 1 \cdots S\}$, and define the partial order $\geqq$ on L (resp. $\mathbb{R}^S$) by $x \geqq y \Leftrightarrow x_i \geqq y_i$, $i = 1 \cdots S$.

LEMMA 1. *L is a complete lattice.*

Proof. Let $A \subset L$. If $A = \varnothing$ then $\inf A = l$, or else the inf if given component-wise.

LEMMA 2. *T is isotone.*

Proof. From $x \geqq y$ it follows that $\sum_{j=1}^S p_{ij}(a)\, x_j \geqq \sum_{j=1}^S p_{ij}(a)\, y_j$, $i = 1 \cdots S$, $a = 1 \cdots A$, since $p_{ij}(a)$ are not negative. Adding $r(i, a)$ and taking the maximum over all a yields $Tx \geqq Ty$.

LEMMA 3. *T maps L to itself.*

Proof. The proof given in [7] is straightforward, using the contraction property of T. An alternative proof is given by the fact that $Tl \leqq l$, $T(-l) \geqq -l$ and by the monotonicity of T (note that l is the 1 and $-l$ the 0 of the lattice L).

From the proof of Birkhoff's theorem it follows that a solution of (OE) can be obtained by determining the least element v of L such that $v \geqq Tv$, that is,

$$V = \min\{v \in L: v \geqq Tv\}. \tag{1}$$

Since the minimum in (1) exists, it can be obtained by minimizing $\sum_{i=1}^S v_i$.

Moreover, $v \geqq Tv$ is equivalent to $v_i \geqq r(i, a) + \beta \sum_{j=1}^{S} p_{ij}(a)\, v_j$, $i = 1 \cdots S$, $a = 1 \cdots A$. Hence, a solution of (OE) is given by the optimal solution of

$$\sum_{i=1}^{S} v_i \to \min,$$

$$v_i - \beta \sum_{j=1}^{S} p_{ij}(a)\, v_j \geqq r(i, a), \qquad i = 1 \cdots S, \quad a = 1 \cdots A,$$

$$|v_i| \leqq \lambda, \qquad i = 1 \cdots S,$$

which is a linear programming problem (LP).

The uniqueness of the optimal value function follows easily from the contraction property of T. Since $v \geqq Tv$ implies $v \geqq V$ for any $v \in \mathbb{R}^S$, the fixed point V of T is also given by

$$V = \min\{v \in \mathbb{R}^S : v \geqq Tv\}, \tag{2}$$

and is the optimal solution to

$$\sum_{i=1}^{S} v_i \to \min,$$

$$v_i - \beta \sum_{j=1}^{S} p_{ij}(a)\, v_j \geqq r(i, a), \qquad i = 1 \cdots S, \quad a = 1 \cdots A,$$

$$v_i \in \mathbb{R}, \qquad i = 1 \cdots S.$$

The (lattice-theoretic) dual to (1) and (2) is given by

$$V = \max\{v \in L : v \leqq Tv\} = \max\{v \in \mathbb{R}^S : v \leqq Tv\}. \tag{3}$$

This cannot lead to a LP, since the set $\{v \in L : v \leqq Tv\}$ is, in general, not convex (see example below).

Introducing for any policy $f \in F := \{g : S \to A\}$ the operator

$$T_f : \quad \mathbb{R}^S \to \mathbb{R}^S$$

$$v \to \left(r(i, f(i)) + \beta \sum_{j=1}^{S} p_{ij}(f(i))\, v_j \right), \qquad i = 1 \cdots S$$

and the value function V_f as the fixed point of T_f, we have as above

$$V_f = \max\{v \in \mathbb{R}^S : v \leqq T_f v\}.$$

Since $Tv = \max\{T_f v, f \in F\}$, we can write (3) as

$$V = \max\{v: \exists f \in F: v \leqq T_f v\} = \max \bigcup_{f \in F} \{v: v \leqq T_f v\}$$
$$= \max\{\max\{v: v \leqq T_f v\}, f \in F\} = \max\{V_f, f \in F\},$$

hence V is the maximum over all value functions V_f, $f \in F$.

EXAMPLE. Let $A = S = 2$, $r(i, a) = a - 1$, $a = 1, 2$, $i = 1, 2$;

$$\begin{aligned} p_{ij}(a) &= 1, \qquad i+j+a \text{ odd}, \\ &= 0, \qquad i+j+a \text{ even}, \end{aligned} \qquad \text{and } \beta < 1.$$

Then $\{v \in \mathbb{R}^S: v \leqq Tv\} = \{v \in \mathbb{R}^2: v_1 \leqq 1 + \beta v_2,\ v_2 \leqq 1 + \beta v_1\} \cup \{v \in \mathbb{R}^2: v_1 \leqq 0,\ v_2 \leqq 0\}$, which is not a convex set, while $\{v \in \mathbb{R}^S: v \geqq Tv\} = \{v \in \mathbb{R}^2: v_1 \geqq 1 + \beta v_2,\ v_2 \geqq 1 + \beta v_1\}$ is convex.

REFERENCES

[1] S. M. ROSS, "Introduction to Stochastic Dynamic Programming," Academic Press, New York, 1983.

[2] H. MINE AND S. OSAKI, "Markovian Decision Processes," Amer. Elsevier, New York, 1970.

[3] E. V. DENARDO, Contraction mappings in the theory underlying dynamic programming, *SIAM Rev.* **9** (1967), 165–177.

[4] K. HINDERER, "Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter," Springer–Verlag, Berlin, 1970.

[5] J. F. SHAPIRO, Turnpike planning horizons for a Markovian decision model, *Management Sci.* **14** (1968), 292–300.

[6] K. HINDERER AND G. HÜBNER, An improvement of J. F. Shapiro's turnpike theorem for the horizon of finite stage discrete dynamic programs, *in* "Transactions of the Seventh Praque Conference on Information Theory, Stat. Decision Fct., Random Processes and of the 1974 European Meeting of Statisticians," pp. 245–255, Reidel, Dordrecht, 1974.

[7] J. F. SHAPIRO, Brower's fixed point theorem and finite state Markov decision theory, *J. Math. Anal. Appl.* **49** (1975), 710–712.

[8] H. SCARF, The approximation of fixed points of a continuous mapping, *SIAM J. Appl. Mat.* **15** (1967), 1328–1343.

[9] F. D'EPENOUX', Sur un problème de production et de Stockage dans l'aléatoire, *Rev. Fr. Rech. Oper.* **14** (1960).

[10] A. S. MANNE, Linear programming and sequential decisions, *Management Sci.* **6** (1960), 259–267.

[11] L. C. M. KALLENBERG, "Linear Programming and Finite Markov Control Problems," Mathematisch Centrum, Amsterdam, 1980.

[12] G. BIRKHOFF, "Lattice Theory," Amer. Math. Soc. Providence, R.I., 1967.

[13] A. TARSKI, A lattice-theoretic fixpoint theorem and its applications, *Pacific J. Math.* **5** (1955), 285–309.