RAIRO. RECHERCHE OPÉRATIONNELLE

PAUL J. SCHWEITZER

Solving MDP functional equations by lexicographic optimization

RAIRO. Recherche opérationnelle, tome 16, n° 2 (1982), p. 91-98

<http://www.numdam.org/item?id=RO_1982__16_2_91_0>

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 16, n° 2, mai 1982, p. 91 à 98)

SOLVING MDP FUNCTIONAL EQUATIONS BY LEXICOGRAPHIC OPTIMIZATION (*)

by Paul J. SCHWEITZER (1)

Abstract. — *A vector which lexicographically maximizes the components of the return vector is shown to satisfy the functional equations of infinite horizon discrete dynamic programming.*

Keywords: Markovian decision process; functional equations; lexicographic optimization.

Résumé. — *On montre qu'un vecteur qui maximise lexicographiquement les composants du vecteur des revenus vérifie les équations fonctionnelles de la programmation dynamique discrète à horizon infini.*

1. INTRODUCTION

Consider a discounted [10, 11] or undiscounted [5, 10, 11] semi-Markovian decision process (MDP) in the stationary infinite-horizon setting. A central issue is establishing existence of a policy which is optimal in every state: a policy which *simultaneously* maximizes every component of the return vector (cumulative discounted reward vector or gain rate vector, respectively). A direct proof is given in [17]. The simplest existence proof consists of establishing existence of a solution to the MDP *functional equations* (*see* below), and then showing that any policy which simultaneously achieves all maxima in these functional equations also maximizes all components of the return vector.

Establishing solvability of the functional equations is simplest in the discounted case, where existence of a fixed point to a *contraction operator* (or n-step contraction operator) is always guaranteed [4]. This approach fails in the undiscounted case, where the desired fixed point is of a *non-contractive* operator. Several methods have been proposed for this case, and enumerated in [8]. The earliest method, for both undiscounted and discounted cases, is the policy iteration algorithm (PIA) [5, 10, 11, 19] which generates a sequence of return vectors having finite convergence (if the state and policy spaces are finite) to the maximum-return vector.

(*) Received May 1981.
(1) The Graduate School of Management, The University of Rochester, Rochester, NY 14627, U.S.A.

R.A.I.R.O. Recherche opérationnelle/Operations Research, 0399-0559/1982/91 /$ 5.00

The goal of this paper is to provide a concise alternate proof of the solvability of the functional equations. It proceeds by initiating the PIA with a vector which *lexicographically optimizes* all components of the return vector. The PIA is shown to have immediate convergence to this vector, which implies both that this vector satisfies the functional equations and that some policy simultaneously maximizes all components of the return vector. The new proof is of interest for two reasons, aside from its simplicity: it provides an alternate characterization of the maximal return vector, and it circumvents the technical issues of convergence of the PIA which arise when there are infinitely many states or policies.

2. DISCOUNTED MDP's

The discounted case is presented first, due to its greater simplicity. The functional equations to be solved are:

$$v_i^* = \max_{k \in K(i)} [q_i^k + \sum_{j \in \Omega} M_{ij}^k v_j^*], \qquad i \in \Omega, \tag{1}$$

where $\Omega \equiv \{1, 2, 3, \ldots\}$ denotes the finite or denumerable state-space, $K(i)$ denotes the action space in state i, $K \equiv \underset{i \in \Omega}{X} K(i)$ is the policy space, and q_i^k and M_{ij}^k are the one-step expected reward and discounted transition probability to state j if action k is chosen in state i. These satisfy:

$$|q_i^k| \leqq A < \infty, \qquad M_{ij}^k \geqq 0, \qquad \sum_{j \in \Omega} M_{ij}^k \leqq \beta < 1.$$

For each stationary policy $f = [f(1), f(2), f(3), \ldots] \in K$, where $f(i) \in K(i)$ is the action used in state i, define:

$$q(f) = [q(f)_i] = [q_i^{f(i)}], \qquad M(f) = [M(f)_{ij}] = [M_{ij}^{f(i)}],$$

and the return vector:

$$v(f) \equiv [I - M(f)]^{-1} q(f) \equiv \sum_{n=0}^{\infty} M(f)^n q(f), \tag{2}$$

which is the unique bounded solution ($|v(f)|_\infty \leqq A/(1-\beta)$) to the equations:

$$v(f) = q(f) + M(f)\, v(f). \tag{3}$$

We require that K be compact and that $v(f)$ be continuous on K. These are met automatically if the state space Ω and every action space $K(i)$ is finite, which is the situation when numerical computations are undertaken. If Ω is denumerable or any $K(i)$ is non-finite, additional technical assumptions are needed to meet these requirements: if each $K(i)$ is compact, then Tychonoff's theorem [12] ensures the compactness of K, while the continuity of $q(f)$ and $M(f)$ ensures the continuity of $v(f)$.

Our goal is to find a policy $f^* \in K$ such that $v(f^*)$ solves equation (1). It is then straightforward to show [2] that:

$$v(f^*)_i = \max_{f \in K} v(f)_i, \qquad i \in \Omega, \tag{4}$$

so that f^* simultaneously achieves all maxima on the rightside of equation (4) and is optimal in every state. Such a policy is obtained as follows.

Define $\{\bar{v}_i\}_{i=1}^{\infty}$ and sets $\{S_i\}_{i=0}^{\infty} \subseteq K$ recursively by:

$$S_0 \equiv K,$$

$$\bar{v}_i \equiv \max\{v(f)_i \mid f \in S_{i-1}\} = \max\{v(f)_i \mid f \in K \text{ and } v(f)_j = \bar{v}_j \text{ for } j < i\},$$

$$i = 1, 2, 3, \ldots,$$

$$S_i \equiv \{f \in S_{i-1} \mid v(f)_i = \bar{v}_i\} = \{f \in K \mid v(f)_j = \bar{v}_j \text{ for } j \leqq i\},$$

$$i = 1, 2, 3, \ldots,$$

Note that the vector $[\bar{v} = \bar{v}_1, \bar{v}_2, \bar{v}_3, \ldots]$ lexicographically maximizes all components of the vector $v(f) = [v(f)_1, v(f)_2, \ldots]$ over $f \in K$, and that:

$$K = S_0 \supseteq S_1 \supseteq S_2 \supseteq \ldots \supseteq S_i \neq \emptyset, \qquad i \geqq 1.$$

If K is compact and $v(f)$ is continuous on K, it is claimed that, for any policy $f^* \in \bigcap_{i \geqq 1} S_i$ (this intersection is non-empty by a modification of Cantor's intersection theorem [1, p. 77]) $v(f^*) = \bar{v}$ satisfies (1).

To show this, note first that the expression defining $\bar{v}_1$ is the maximum of a continuous function on a compact set, hence is achieved. Therefore S_1 is non-empty and is also compact. Inductively, each $\bar{v}_i$ is well-defined and each S_i is non-empty and compact.

Start the PIA [10, 11] with $v(f^*)$ and do one policy improvement step. This determines a new policy h with either (*a*) $h = f^*$ and $v(f^*)$ solves the functional equations (1), or (*b*) $v(h)_i \geqq v(f^*)_i$ for every i with strict inequality for at least one i.

To show that case (*b*) cannot occur, note that:

$$\bar{v}_1 \equiv \max_{f \in K} v(f)_1 \geqq v(h)_1 \geqq v(f^*)_1 = \bar{v}_1,$$

so that $h \in S_1$. Then:

$$\bar{v}_2 \equiv \max_{f \in S_1} v(f)_2 \geqq v(h)_2 \geqq v(f^*)_2 = \bar{v}_2,$$

so that $h \in S_2$. Proceeding inductively, $v(f^*)_i = \bar{v}_i = v(h)_i$ for every i. Case (*a*) shows $v(f^*)$ satisfies (1).

3. UNDISCOUNTED CASE

The procedure is illustrated for the case of 3 nested functional equations, which arise when seeking *maximum-bias policies* [6, 7, 14, 19]. By discarding the last equation, the procedure reduces to one for finding *maximal-gain policies*. A straightforward extension to four or more nested functional equations will generalize the procedure to higher-order optimality criteria [7, 14, 20].

The nested functional equations to be solved are:

$$g_i^* = \max_{k \in K(i)} [\sum_{j \in \Omega} p_{ij}^k g_j^*], \qquad i \in \Omega, \tag{5a}$$

$$w_i^* = \max_{k \in L(g^*, i)} [q_i^k - \sum_{j \in \Omega} H_{ij}^k g_j^* + \sum_{j \in \Omega} P_{ij}^k w_j^*], \quad i \in \Omega, \tag{5b}$$

$$y_i^* = \max_{k \in M(g^*, w^*, i)} [a_i^k + \sum_{j \in \Omega} B_{ij}^k g_j^* - \sum_{j \in \Omega} H_{ij}^k w_j^* + \sum_{j \in \Omega} P_{ij}^k y_j^*], \qquad i \in \Omega, \tag{5c}$$

where:

$$L(g^*, i) \equiv \{k \in K(i) \mid g_i^* = \sum_{j \in \Omega} P_{ij}^k g_j^*\}, \qquad i \in \Omega,$$

$$M(g^*, w^*, i) \equiv \{k \in L(g^*, i) \mid w_i^* = q_i^k - \sum_{j \in \Omega} H_{ij}^k g_j^* + \sum_{j \in \Omega} P_{ij}^k w_j^*\}, \qquad i \in \Omega,$$

$$P_{ij}^k \geqq 0, \qquad \sum_{j \in \Omega} P_{ij}^k = 1,$$

$$H_{ij}^k \geqq 0, \qquad \sum_{j \in \Omega} H_{ij}^k \leqq A_1 < \infty, \qquad H_{ij}^k = 0 \quad \text{if } P_{ij}^k = 0,$$

$$B_{ij}^k \geqq 0, \qquad \sum_{j \in \Omega} B_{ij}^k \leqq A_2 < \infty, \qquad B_{ij}^k = 0 \quad \text{if } H_{ij}^k = 0,$$

$$|q_i^k| \leqq A_3 < \infty, \qquad |a_i^k| \leqq A_4 < \infty.$$

For any policy $f \in K \equiv \underset{i \in \Omega}{X} K(i)$, the equations:

$$\left.\begin{aligned} g(f) &= P(f)g(f), \qquad w(f) = q(f) - H(f)g(f) + P(f)w(f), \\ y(f) &= a(f) + B(f)g(f) - H(f)w(f) + P(f)y(f), \end{aligned}\right\} \quad (6a, b, c)$$

will have a unique solution set $\{g(f), w(f), y(f)\}$ provided supplementary conditions are supplied, one per subchain of $P(f)$, to fix the arbitrary additive constants in $y(f)$. We require that K be compact and that $g(f)$, $w(f)$, $y(f)$ be continuous on K. These are met automatically if the state and action spaces are finite. If not, additional technical assumptions are needed to meet these requirements, including (*a*) $P(f)$, $H(f)$, $B(f)$, $q(f)$, $a(f)$ are continuous in f; (*b*) every $T_i^k \equiv \sum_{j \in \Omega} H_{ij}^k \geqq \varepsilon_1 > 0$ (to ensure $g(f)$ is bounded); (*c*) for every policy $f \in K$, $P(f)$ has the same number of subchains (to ensure [16] that $P^*(f) =$ Cesaro-limit of $P(f)^n$, and therefore $g(f)$, are continuous in f); (*d*) if $P_{ij}^k > 0$, then $P_{ij}^k \geqq \varepsilon > 0$ (to ensure that the fundamental matrix of $P(f)$, and therefore $w(f)$, is bounded). (*a*, *b*, *c*) are similar to the assumptions that $g(f)$, $P(f)$ and $P^*(f)$ are each continuous in $f \in K$, used by Sheu and Farn [18, thm. 3.3] to ensure existence of a stationary 1-optimal policy for a MDP with finite state space and compact action space.

Define $\{\bar{g}_i, \bar{w}_i, \bar{y}_i\}_{i=1}^{\infty}$ and sets $\{S_i, U_i, Z_i\}_{i=0}^{\infty}$ recursively by:

$$S_0 \equiv K,$$

$$\bar{g}_i \equiv \max_{f \in S_{i-1}} g(f)_i, \qquad i = 1, 2, 3, \ldots,$$

$$S_i \equiv \{f \in S_{i-1} \mid g(f)_i = \bar{g}_i\}, \qquad i = 1, 2, 3, \ldots$$

Note that any policy in $U_0 \equiv \bigcap_{i \geqq 1} S_i$ lexicographically maximizes all components of $g(f) = [g(f)_1, g(f)_2, g(f)_3, \ldots]$ over $f \in K$.

$$\bar{w}_i \equiv \max_{f \in U_{i-1}} w(f)_i, \qquad i = 1, 2, 3, \ldots,$$

$$U_i = \{f \in U_{i-1} \mid w(f)_i = \bar{w}_i\}, \qquad i = 1, 2, 3, \ldots$$

Note that any policy in $Z_0 \equiv \bigcap_{i \geqq 1} U_i$ lexicographically maximizes all components of $w(f) = [w(f)_1, w(f)_2, w(f)_3, \ldots]$ over $f \in U_0$.

$$\bar{y}_i \equiv \max_{f \in Z_{i-1}} y(f)_i, \qquad i = 1, 2, 3, \ldots,$$

$$Z_i \equiv \{f \in Z_{i-1} \mid y(f)_i = \bar{y}_i\}, \qquad i = 1, 2, 3, \ldots$$

Note that any policy in $Z_\infty \equiv \bigcap_{u \geq 1} Z_i$ lexicographically maximizes all components of $y(f)=[y(f)_1, y(f)_2, y(f)_3, \ldots]$ over $f \in Z_0$. Formally, any policy in Z_∞ lexicographically maximizes all components of $\{g(f), w(f), y(f)\}$, with outcome $\{\bar{g}, \bar{w}, \bar{y}\}$. Note also the nesting property:

$$K \supseteq S_1 \supseteq S_2 \supseteq \ldots \supseteq U_0 \supseteq U_1 \supseteq U_2 \supseteq \ldots \supseteq Z_0 \supseteq Z_1 \supseteq Z_2 \ldots \supseteq Z_\infty \neq \emptyset.$$

The procedure is well-defined if K is compact and $g(f)$, $w(f)$, $y(f)$ are continuous on K: all maxima will exist and each S_i, U_i, Z_i is compact.

For any policy $f^* \in Z_\infty$, it is claimed that the triple $\{\bar{g}, \bar{w}, \bar{y}\} = \{g(f^*), w(f^*), y(f^*)\}$ satisfies the functional equations (5 *a*, *b*, *c*). It is then straightforward [15, thm. 6.17] to show that:

$$g(f^*)_i = \max_{f \in K} g(f)_i, \qquad i \in \Omega,$$

$$w(f^*)_i = \max \{w(f)_i \mid f \in K \text{ with } g(f)=g(f^*)\}, \qquad i \in \Omega.$$

To establish the claim, enter the PIA for these 3 equations [14, 19] with the initial vector triple $\{\bar{g}, \bar{w}, \bar{y}\} = \{g(f^*), w(f^*), y(f^*)\}$, and let one policy improvement step produce a successor policy h. Then one of 4 cases holds:

(i) $h=f^*$ and $\{g(f^*), w(f^*), y(f^*)\}$ solve (5),

(ii) $g(h) \geq g(f^*)$ and $g(h) \neq g(f^*)$,

(iii) $g(h)=g(f^*)$, $w(h) \geq w(f^*)$, and $w(h) \neq w(f^*)$,

(iv) $\begin{cases} g(h)=g(f^*), & w(h)=w(f^*), \\ y(f) \geq y(f^*), & \text{and} \quad y(f) \neq y(f^*). \end{cases}$

Case (ii) is impossible *via* the same reasoning used in the discounted case to show that $v(h)=v(f^*)$. The same reasoning then shows case (iii) is impossible and that case (iv) is impossible. The remaining case (i) confirms the claim.

4. GENERALIZATION

The general structure encompassing the above examples involves the functional equations:

$$x_i^* = \max_{k \in A(x^*, i)} F_i(x^*, k), \qquad i \in S, \tag{7}$$

with $F_i(x, k)$ a given continuous scalar function with arguments $i \in S$ = state space, $k \in K(i)$ = action space in state i, and $x=[x_j]_{j \in S}$ = return vector. The given function $A(x, i)$ is a non-empty set in $K(i)$. The two requirements we impose are:

Policy evaluation step: For each $f=[f(1), f(2), \ldots] \in K \equiv \underset{i \in S}{X} K(i)$ there exists a unique $x(f)=[x(f)_j]_{j \in S}$ satisfying:

$$x(f)_i = F_i(x(f), f(i)), \qquad i \in S. \tag{8}$$

Furthermore, $f(i) \in A(x(f), i)$, $i \in S$.

Policy improvement step: For any policy $f \in K$, define a successor policy $h=[h(1), h(2), \ldots]$ where $h(i) \in A(x(f), i)$ achieves $\max_{k \in A(x(f), i)} F_i(x(f), k)$ and $h(i)=f(i)$ whenever possible. Then [2] either $x(h) \geqq x(f)$ lexicographically, with $x(h) \neq x(f)$, or else $h=f$, and $x^*=x(f)$ satisfies the functional equation (7).

Then with appropriate compactness and continuity assumptions on K and $x(f)$, there exists a policy $f^* \in K$ such that $x(f)$ solves (7). Namely, choose any policy which achieves all maxima in:

$$\bar{x}_1 \equiv \max_{f \in K} x(f)_1,$$

$$\bar{x}_i \equiv \max \{x(f)_i \mid f \in K \text{ with } x(f)_j = \bar{x}_j \text{ for } j < i\}, \qquad i \geqq 2,$$

and furthermore $x(f^*)=\bar{x}$.

To illustrate how the three coupled functional equations in Section 3 fit into this framework: take $S \equiv \{1, 2, 3\} \times \Omega$. For $i=(l, j) \in S$, take $x_{(l, j)}=g_j$ if $l=1$, w_j if $l=2$, and y_j if $l=3$. For $l=1$, $x^*_{(1, j)}=g^*_j$, $A(x, i)=K(j)$, (7) is (5 *a*), and (8) is (6 *a*). For $l=2$, $x^*_{(2, j)}=w^*_j$, $A(x, i)=\{k \in K(j)$ achieving $\max_{k \in K(j)} \sum_{t \in \Omega} P^k_{jt} x_{(1, t)}\}$ with $A(x^*, i)=L(g^*, j)$, (7) *is* (5 *b*), *and* (8) *is* (6 *b*). For $l=3$,

$$x^*_{(l, j)}=y^*_j,$$

$$A(x, i)=\{k \in K(j) \text{ achieving } \max_{k \in K(j)} \sum_{t \in \Omega} P^k_{jt} x_{(1, t)}$$

with ties broken by achieving

$$\max_k \{q^k_j - \sum_{t \in \Omega} H^k_{jt} x_{(2, t)} + \sum_{t \in \Omega} P^k_{jt} x_{(3, t)}\} \text{ with } A(x^*, i)=M(g^*, w^*, j); \tag{7}$$

is (5 *c*) and (8) is (6 *c*).

This generalization provides a framework for understanding the PIA's appearing in the generalized MDP's [3], Leontief substitution systems [9, 13], and n coupled functional equations arising in higher-order optimality conditions [7, 14, 19, 20].

ACKNOWLEDGEMENTS

I thank Marshall Freimer for helpful discussions.

REFERENCES

1. R. G. BARTLE, *The Elements of Real Analysis*, Wiley, New York, second edition, 1976.
2. R. BELLMAN, *Functional Equations in the Theory of Dynamic Programming V. Positivity and Quasi-Linearity*, Proc. Nat. Acad. Sc. U.S.A., Vol. 41, 1955, pp. 743-746.
3. I. BROSH, E. SHLIFER and P. SCHWEITZER, *Generalized Markovian Decision Processes*, Zeitschrift fur Operations Research, Vol. 21, 1977, pp. 173-186.
4. E. DENARDO, *Contraction Mappings in the Theory Underlying Dynamic Programming*, S.I.A.M. Rev., Vol. 9, 1967, pp. 165-177.
5. E. DENARDO and B. FOX, *Multichain Markov Renewal Programs*, S.I.A.M. J. Appl. Math., Vol. 16, 1968, pp. 468-487.
6. E. V. DENARDO, *Computing a Bias-optimal Policy in a Discrete-time Markov Decision Problem*, Oper. Res., Vol. 18, 1970, pp. 279-289.
7. E. V. DENARDO, *Markov Renewal Programs with Small Interest Rates*, Ann. Math. Statist., Vol. 42, 1971, pp. 477-496.
8. A. FEDERGRUEN and P. J. SCHWEITZER, *A Fixed Point Approach to Undiscounted Markov Renewal Programs*, Working Paper 8024, Graduate School of Management, University of Rochester, Rochester, New York, 1980; Also Columbia University, Graduate School of Business Working Paper 351 A, 1980 (Revised 1981).
9. R. C. GRINOLD, *A Generalized Discrete Dynamic Programming Model*, Management Science, Vol. 20, 1974, pp. 1092-1103.
10. R. A. HOWARD, *Dynamic Programming and Markov Processes*, Wiley, New York, 1960.
11. W. JEWELL, *Markov Renewal Programming*, Operations Research, Vol. 11, 1963, pp. 938-971.
12. J. L. KELLEY, *General Topology*, Van Nostrand, Princeton, New Jersey, 1955.
13. G. J. KOEHLER, A. B. WHINSTON and G. P. WRIGHT, *Optimization Over Leontief Substitution Systems*, North-Holland, Amsterdam, 1975.
14. B. L. MILLER and A. F. VEINOTT Jr., *Discrete Dynamic Programming with a Small Interest Rate*, Ann. Math. Statist., Vol. 40, 1969, 366-370.
15. S. M. ROSS, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.
16. P. SCHWEITZER, *Perturbation Theory and Finite Markov Chains*, J. Appl. Prob., Vol. 5, 1968, pp. 401-413.
17. P. J. SCHWEITZER and B. GAVISH, *An Optimality Principle for Markovian Decision Processes*, J. Math. Anal. and Appl., Vol. 54, 1976, pp. 173-184.
18. S. S. SHEU and K.-J. FARN, *A Sufficient Condition for the Existence of a Stationary 1-Optimal Plan in Compact Action Markovian Decision Processes*. Recent Developments in Markov Decision Processes, R. HARTLEY, L. C. THOMAS, D. J. WHITE, Eds., Academic Press, New York, 1980, pp. 111-126.
19. A. F. Jr. VEINOTT, *On Finding Optimal Policies in Discrete Dynamic Programming with No Discounting*, Ann. Math. Statist., Vol. 37, 1966, pp. 1284-1294.
20. A. F. Jr. VEINOTT, *Discrete Dynamic Programming with Sensitive Discount Optimality Criteria*, Ann. Math. Statist., Vol. 40, 1969, pp. 1635-1660.