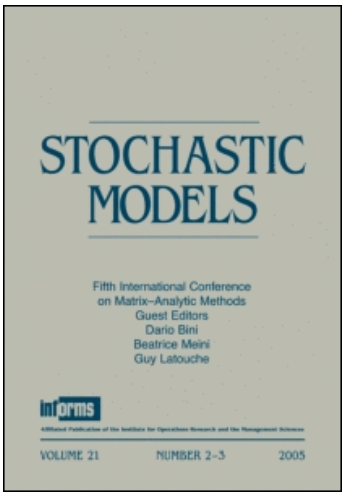


Stochastic Models
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597301

Counter examples for compact action markov decision chains with average reward criteria

Rommert Dekker [a]
[a] University of Leiden,

Online Publication Date: 01 January 1987

To cite this Article Dekker, Rommert(1987)'Counter examples for compact action markov decision chains with average reward criteria',Stochastic Models,3:3,357 — 368
To link to this Article: DOI: 10.1080/15326348708807061
URL: http://dx.doi.org/10.1080/15326348708807061

COMMUN. STATIST.-STOCHASTIC MODELS, 3(3), 357-368 (1987)

COUNTER EXAMPLES FOR COMPACT ACTION MARKOV DECISION CHAINS WITH AVERAGE REWARD CRITERIA

Rommert Dekker[1]

University of Leiden

ABSTRACT

In this note we present two examples of compact-action finite-state Markov decision chains in which a policy improvement procedure yields wrong or limited results. In the first example, which exhibits a multichain structure, there is no convergence of the average rewards of the successive policies to the maximal value. In the second example, which has a unichain structure, the lack of uniqueness of maximizing policies in each step of the algorithm means that there is no convergence of either bias vectors or maximizing policies. Accordingly, no solution to the average optimality equations can be obtained.

1. INTRODUCTION.

The policy improvement procedure (PIP) is a well-established optimization technique in finite-state, finite-action Markov decision chains, with respect to both average and discounted reward criterion. Since in each iteration of the procedure the objective function increases, the finiteness of the number of stationary and deterministic policies (together with an anticycling rule) guarantees that the algorithm will end with an optimal policy.

---

1) Research was sponsored by the Netherlands Foundation for Mathematics (SMC). Present address: Koninklijke/Shell Laboratorium Amsterdam, P.O. Box 3003, 1003 AA Amsterdam.

Recently, investigations have been carried out into the convergence of policy improvement in finite-state compact-action Markov decision chains ([1],[2],[3]). A relationship with Newton's method proved to be quite successful and for discount optimality convergence of the successive policies to the optimum policy could be shown straightforwardly ([1]). The analysis for the average reward criterion, however, appeared to be more complicated ([2] and [3]). Recall that in a compact-action Markov decision chain conditions are required for the existence of solutions to the average optimality equations. Apart from the basic assumption of continuity of transition probabilities and immediate rewards in the action, the continuity of the number of minimal closed sets is a sufficient condition (see [4]). However, it was not clear whether these conditions also guarantee the success of policy improvement. In Hordijk and Puterman [2], convergence of the iterates to a solution to the average optimality equations was shown only for the unichain case under uniqueness of the maximizing policy. In this note, which originates from [3], we shall show by counter examples that both the assumptions made in [2] are necessary.

2. THE MODEL.

Let E denote the finite state space and A(i) the compact set of available actions in state i. Given action a in state i, a transition is made to state j with probability $P_{ij}(a)$, and an immediate reward $r_i(a)$ is obtained. Both $P_{ij}(a)$ and $r_i(a)$ are continuous in action a for all $i,j \in E$. In this paper we only consider deterministic, stationary Markov policies. Given such a policy f we denote the corresponding matrix of transition probabilities by P(f). We speak of the unichain case if under each policy the induced Markov chain has one minimal closed set and else of the multichain case. Let $\Pi(f)$ denote the stationary matrix, i.e. the Cesaro limit of $P^k(f)$ for $k \to \infty$. For policy f, its average reward g(f) and relative value vector v(f) can be obtained from the set of equations

$$\left.\begin{aligned} g &= P(f)g \\ g + v &= r(f) + P(f)v \end{aligned}\right\} \quad (1)$$

The vector g is uniquely determined by equations (1), the vector v, however, is not. The relative value vector can be determined uniquely by adding the equation

$$\Pi(f)v = 0, \quad (2)$$

which also serves as an anticycling rule in the context of a policy improvement procedure.

3. THE POLICY IMPROVEMENT PROCEDURE FOR THE AVERAGE REWARD.

Following [2] we shall formulate the policy improvement procedure through operators $B^{(-1)}$, $B^{(0)}$ and nested action sets $A^{(-1)}(i)$ and $A^{(0)}(i)$ defined by

$$B_i^{(-1)}(g^{(n)}) \equiv \max_{a \in A(i)} \sum_j P_{ij}(a) g_j^{(n)} - g_i^{(n)}, \quad i \in E, \quad (3)$$

$$A^{(-1)}(i) \equiv \{ i \in A(i) \mid \sum_j P_{ij}(a) g_j^{(n)} - g_i^{(n)} = B_i^{(-1)}(g^{(n)}) \}$$

$$B_i^{(0)}(g^{(n)}, v^{(n)}) \equiv \max_{a \in A^{(-1)}(i)} \left[r_i(a) + \sum_j P_{ij}(a) v_j^{(n)} - v_i^{(n)} - g_i^{(n)} \right], \quad i \in E \quad (4)$$

$$A^{(0)}(i) \equiv \{ a \in A^{(-1)}(i) \mid r_i(a) + \sum_j P_{ij}(a) v_j^{(n)} - v_i^{(n)} - g_i^{(n)} = B_i^{(0)}(g^{(n)}, v^{(n)}) \}$$

Interpret $A^{(-1)}(i)$ as the set of actions achieving the maximum in (3) and $A^{(0)}(i)$ similarly in (4). The policy improvement procedure can now be written as follows ($g(f^{(n)})$, $v(f^{(n)})$ will be abbreviated by $g^{(n)}, v^{(n)}$ resp.).

step 1- initialization: choose any $f^{(1)} \in F$, set $n = 1$.

step 2- policy evaluation: evaluate $g^{(n)}$, $v^{(n)}$ from equations (1) and (2).

step 3- policy improvement: determine for all $i \in E$ $B_i^{(-1)}(g^{(n)}, v^{(n)})$ $B_i^{(0)}(g^{(n)}, v^{(n)})$ and the corresponding sets of maximizing actions $A^{(-1)}(i)$, $A^{(0)}(i)$. Choose $f^{(n+1)}(i)$ from $A^{(0)}(i)$, if possible choose $f^{(n+1)}(i)=f^{(n)}(i)$. If $f^{(n+1)}(i)=f^{(n)}(i)$ for all $i \in E$ then stop, else let n:=n+1 and go to step 2.

4. CONVERGENCE OF POLICY IMPROVEMENT IN THE GENERAL CASE.

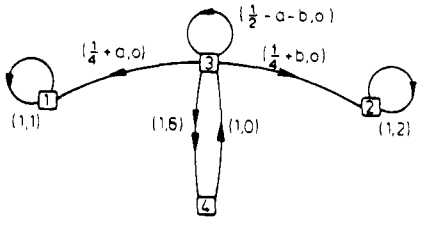
Similarly to the finite action case, it can be shown that in each step of the algorithm for each state $i \in E$ either

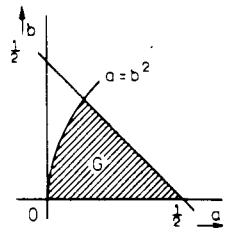
$$\left.\begin{array}{ll} g_i^{(n+1)} > g_i^{(n)} & \text{or} \\ g_i^{(n+1)} = g_i^{(n)} & \text{and} \quad v_i^{(n+1)} \geq v_i^{(n)}. \end{array}\right\} \qquad (5)$$

If $g_i^{(n+1)} = g_i^{(n)}$ and $v_i^{(n+1)} = v_i^{(n)}$ for all $i \in E$, then the algorithm has stopped as it has obtained a solution to the average optimality equations. Since action sets are compact rather than finite the algorithm is not necessarily finite. As it can easily be shown that g(f) is bounded from above, it follows from (5) that the iterates $g^{(n)}$ converge, though not always to the maximal value as a the following theorem states.

THEOREM 1. If there are policies under which the induced Markov chain has a multichain structure then the sequence $g^{(n)}$ generated by the policy improvement procedure does not necessarily converge to the optimal average reward g^*.

PROOF. Consider the following counter example (suggested by Arie Hordijk) with a multichain structure. Let $\boxed{i}$ denotes state i and (x,y) on an arc denotes the transition probability and the reward respectively. In states 1,2 and 4 there is one action only, while in state 3 the set of actions A(3) is: $\{2\} \cup \{(a,b) \mid (a,b) \in G\}$, where action 2 corresponds to the double arrow





and actions (a,b) to the single. The set G is bounded by the curves: $b \geq 0$, $a=b^2$, $a+b=\frac{1}{2}$. Notice that state 1 and 2 are always absorbing and have average reward $g_1=1$ and $g_2=2$. Given action (a,b) in state 3, we can easily compute that $1 < g_3(a,b)=g_4(a,b)=1+\left(\frac{b+\frac{1}{4}}{a+b+\frac{1}{2}}\right) < 2$. For action 2 in state 3 we have $g_3(2)=g_4(2)=3$, which is therefore the average-optimal action. Given action $(a,b) \notin G$ in state 3, what does policy improvement yield? Action 2 in state 3 yields no improvement with respect to $B_3^{(-1)}(g)$, hence we determine

$$\max_{(a,b)\notin G} \left[\sum_j P_{3j}(a,b)g_j - g_3\right] = \max_{(a,b)\in G} \left[(a+\tfrac{1}{4}) + 2(b+\tfrac{1}{4}) + (\tfrac{1}{2}-a-b)g_3\right]$$

$$= \max_{(a,b)\in G} \left[(1-g_3)a + (2-g_3)b + (\tfrac{3}{4} - \tfrac{1}{2}g_3)\right].$$

This is a maximization of a linear function over a convex area G. Hence the maximum is attained at the boundary, and from the coefficients of a and b we see that it is on the curve $a=b^2$, with $0 \leq b \leq -\frac{1}{2} + \frac{1}{2}\sqrt{3}$. Insertion of this yields the unique maximizing action: $(\tilde{b}^2,\tilde{b})$, with

$$\tilde{b} = \min\left(-\tfrac{1}{2} + \tfrac{1}{2}\sqrt{3}, \frac{b^2 + \frac{1}{4}}{2(b + \frac{1}{4})}\right). \tag{6}$$

For any action (b^2,b) with $0 \leq b \leq -\frac{1}{2} + \frac{1}{2}\sqrt{3}$ we find that $B^{(-1)}(g(b^2,b)) \geq 0$, with equality only if $b = b_0 = \frac{1}{4}(-1 + \sqrt{5})$. At this point, we choose action 2 through operator $B^{(0)}$. If we start policy improvement with any action $((b_1)^2,b_1)$, with rational b_1, we see that the iterates have b-values which are rational functions of b_1 or $\sqrt{3}$. Hence it will take an infinite number of steps to attain the policy $((b_0)^2,b_0)$ and although the average reward increases, it remains below 2. Thus the $g^{(n)}$-iterates of the policy improvement procedure converge to a submaximal value. □

In mathematical terms, this example shows that the operator $B^{(0)}$ can be discontinuous. It is easily checked that the policy

iterates $(b_n^2, b_n) \to (b_0^2, b_0)$ and that $g^{(n)} \to g^{(0)} = g(b_0^2, b_0)$ and $v^{(n)} \to v^{(0)} = v(b_0^2, b_0)$ for $n \to \infty$. However, (b_0^2, b_0) is not the maximizing action of $g^{(0)}, v^{(0)}$, since action 2 is. This implies that $B^{(0)}(g^{(n)}, v^{(n)})$ does not converge to $B^{(0)}(g^{(0)}, v^{(0)})$.

In Markov chain terminology, the essence of this example is that an extra closed set has to be created for an average optimal policy. The following theorem shows that that when an improvement is made with the operator $B^{(-1)}$ only, an extra closed set cannot be created.

THEOREM 2. If $B_i^{(-1)}(g(f)) > 0$ for some $i \in E$ and $f \in F$, and policy h is such that $P(h)g(f) - g(f) = B^{(-1)}(g(f))$, then i is a transient state under policy h.

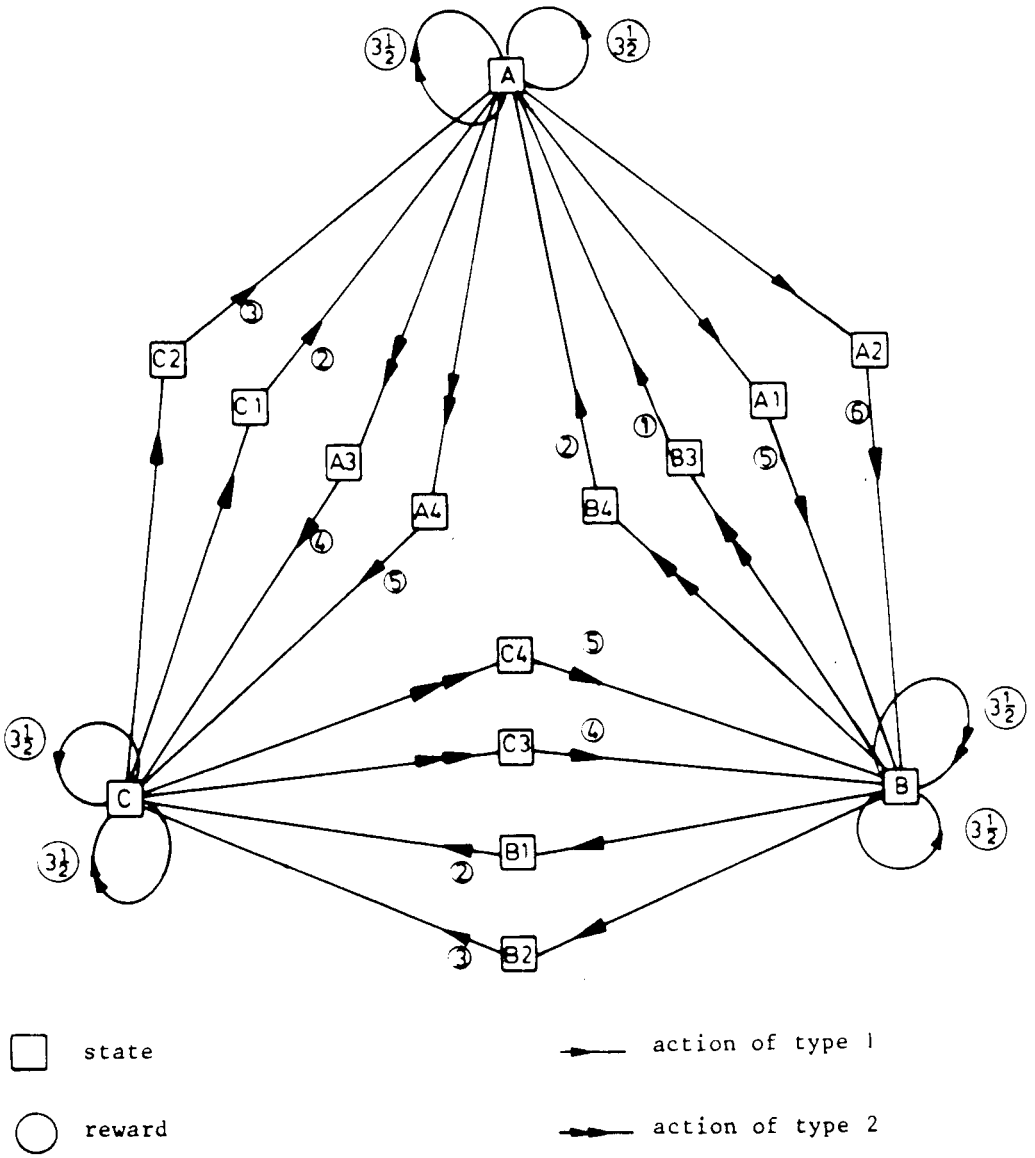
PROOF. Notice that $B^{(-1)}(g(f)) \geq 0$ and that $\Pi(h)B^{(-1)}(g(f)) = 0$, implying that $B_i^{(-1)}(g(f)) = 0$ for all states i recurrent under policy h. □

5. CONVERGENCE OF POLICY IMPROVEMENT IN THE UNICHAIN CASE.

In the unichain case there are no problems of the above type. For this case, Hordijk and Puterman [2] proved that $g^{(n)} \to g^*$, where g^* is the maximal average reward. They also showed convergence of the $v^{(n)}$ iterates to some vector v^*, provided that at each iteration of the algorithm there is a unique maximizing policy. The pair g^*, v^* then constitutes a solution to the average optimality equations. In this case the maximizing policies even converge to the (unique) Blackwell optimal policy. The problem of whether or not the assumption of unique maximizing policies was necessary was an open one. The following theorem, however, states that this assumption is necessary.

THEOREM 3. If there does not exist a unique maximizing policy at each iteration of the PIP then the sequence of policies $f^{(n)}$ and relative value vectors $v^{(n)}$ does not necessarily converge.

PROOF. Consider the following counterexample.
Let E = {A, A1, A2, A3, A4, B, B1, B2, B3, B4, C, C1, C2, C3, C4}
Actions can only be chosen in states A, B and C. These states all have a similar action set: $\{(1,a) \mid 0 \le a \le \frac{1}{2}\} \cup \{(2,a) \mid 0 \le a \le \frac{1}{2}\}$, the actions of the first set are called type 1, those of the second set type 2, the parameter a is called the fraction. The following picture shows the possible transitions and rewards.



The transition probabilities and rewards for state A are:

$$P_{A,A}(1,a) = \frac{3}{4} - a - a^2, \qquad P_{A,A}(2,a) = \frac{3}{4} - a - a^2,$$
$$P_{A,A1}(1,a) = \frac{1}{8} + a^2, \qquad P_{A,A3}(2,a) = \frac{1}{8} + a^2,$$
$$P_{A,A2}(1,a) = \frac{1}{8} + a, \qquad P_{A,A4}(2,a) = \frac{1}{8} + a,$$
$$r_A(1,a) = 3\tfrac{1}{2}, \qquad r_A(2,a) = 3\tfrac{1}{2}.$$

For B and C the same transition probabilities and rewards hold if we replace A by B or C, A1 by B1 or C1, etc. The transition probabilities for the other states are (we state only the non-zero ones and skip the action notation, since only one action is possible):

$$P_{A1,B} = P_{A2,B} = 1, \quad P_{B1,C} = P_{B2,C} = 1, \quad P_{C1,A} = P_{C2,A} = 1.$$
$$P_{A3,C} = P_{A4,C} = 1, \quad P_{B3,A} = P_{B4,A} = 1, \quad P_{C3,B} = P_{C4,B} = 1.$$

The immediate rewards are

$$r_{A1} = 5, \quad r_{B1} = 2, \quad r_{C1} = 2,$$
$$r_{A2} = 6, \quad r_{B2} = 3, \quad r_{C2} = 3,$$
$$r_{A3} = 4, \quad r_{B3} = 1, \quad r_{C3} = 4,$$
$$r_{A4} = 5, \quad r_{B4} = 2, \quad r_{C4} = 5.$$

A policy is determined by the actions chosen in state A, B and C. Note that the Markov chain structure only depends on whether type 1 or type 2 actions are chosen and that there is always a unichain structure. Suppose we start the PIP with a policy which has the same fraction a in each state. First we show that the average reward depend only on the fraction a, and not on the type of action chosen. The same is true for the set of solutions to $g+v = r(f)+P(f)v$. However, the actual value of the bias vector $v(f)$, which is determined by adding the equation $\Pi(f)v=0$, does depend on which types are chosen. For the choice of maximizing actions this is not relevant. Secondly, we shall show that the maximizing policies again have the same fraction in states A, B and C and that in A, B and C we are free to choose type 1 or 2. Finally, we show that it takes in general an infinite number of steps to attain the

maximal average reward. Since in each step we are strictly improving and have the option of taking actions of either type 1 or 2, we do not necessarily have convergence of $v^{(n)}$ or $f^{(n)}$.

First, we calculate the average reward and bias vector for any policy. Regardless of the policy, we always have the following set of equations from $g + v = r(f) + P(f)v$.

$$\left.\begin{array}{lll} g+v_{A1} = 5+v_B, & g+v_{B1} = 2+v_C, & g+v_{C1} = 2+v_A, \\ g+v_{A2} = 6+v_B, & g+v_{B2} = 3+v_C, & g+v_{C2} = 3+v_A, \\ g+v_{A3} = 4+v_C, & g+v_{B3} = 1+v_A, & g+v_{C3} = 4+v_B, \\ g+v_{A4} = 5+v_C, & g+v_{B4} = 2+v_A, & g+v_{C4} = 5+v_B. \end{array}\right\} \quad (7)$$

From equations (7) we see that the values of v in states A1, A2, ... are determined by g and the value of v in A, B and C.
Suppose we have a policy with the same fraction a in each of the states A, B and C. If we choose an action of type 1 in A we find

$$g + v_A = 3\tfrac{1}{2} + (\tfrac{3}{4}-a-a^2)v_A + (\tfrac{1}{8}+a^2)v_{A1} + (\tfrac{1}{8}+a)v_{A2}.$$

By equations (7) we have

$$g + (\tfrac{1}{4}+a+a^2)v_A = 3\tfrac{1}{2} + (\tfrac{1}{8}+a^2)(5-g+v_B) + (\tfrac{1}{8}+a)(6-g+v_B).$$

Hence

$$(1\tfrac{1}{4}+a+a^2)g + (\tfrac{1}{4}+a+a^2)v_A = 5(\tfrac{1}{4}+a+a^2) + (3\tfrac{5}{8}+a) + (\tfrac{1}{4}+a+a^2)v_B.$$

Let

$$\beta(a)=\tfrac{1}{4}+a+a^2, \quad \gamma(a)=(1\tfrac{1}{4}+a+a^2)/\beta(a), \quad \delta(a)=(3\tfrac{5}{8}+a)/\beta(a),$$

then,

$$\gamma(a)g + v_A = v_B + 5 + \delta(a).$$

For the other actions, similar equations are obtained in the same way. Collecting them yields,

action	in state	equation
(1,a)	A	$\gamma(a)g + v_A = v_B + 5 + \delta(a)$
(2,a)	A	$\gamma(a)g + v_A = v_C + 4 + \delta(a)$

(1,a)	B	$\gamma(a)g + v_B = v_C + 2 + \delta(a)$
(2,a)	B	$\gamma(a)g + v_B = v_A + 1 + \delta(a)$
(1,a)	C	$\gamma(a)g + v_C = v_A + 2 + \delta(a)$
(2,a)	C	$\gamma(a)g + v_C = v_B + 4 + \delta(a)$

It is easily verified that for any policy with the same fraction a in A, B and C we have

$$\gamma(a)g = 3 + \delta(a) \text{ and } v_A = v_B + 2,\ v_C = v_B + 1. \tag{8}$$

Evaluating g yields,

$$g = \left[3 + \frac{(3\frac{5}{8}+a)}{\beta(a)}\right] \frac{\beta(a)}{1\frac{1}{4}+a+a^2} = 3 + \frac{\frac{5}{8}+a}{1\frac{1}{4}+a+a^2} \tag{9}$$

We find the actual values of the bias vector by normalizing the v vector with $\Pi(f)v = 0$. Let f_1 take action (1,a) in each state A, B and C. Then all states A, B and C are recurrent and after some calculations we obtain

$$v_C(f_1) = \frac{g(a)}{\gamma(a)} - \frac{1}{\beta(a)\gamma(a)}\left\{\frac{7}{8}+4a+3a^2\right\},$$

from which the other values of v easily follow.
For policy f_2, which differs from policy f_1 only in that it takes action (2,a) instead of action (1,a) in state C, state A becomes transient. After some calculations we find

$$v_C(f_2) = v_C(f_1) + \tfrac{1}{2}.$$

The choice of type does influence the actual v vector!

Secondly, we remark that only the fraction, not the type of the maximizing policy, is uniquely determined. For instance, if we determine the new fraction b, given g, v and old fraction a, then we have to determine for state A and type 1 actions,

$$\max_{0\le b\le \frac{1}{2}} \{3\tfrac{1}{2} + (\tfrac{3}{4}-b-b^2)v_A + (\tfrac{1}{8}+b^2)v_{A1} + (\tfrac{1}{8}+b)v_{A2} - v_A - g\}$$

$$= \max_{0\le b\le \frac{1}{2}} \{3\tfrac{1}{2} - \beta(b)v_A + (\tfrac{1}{8}+b^2)(5-g+v_B) + (\tfrac{1}{8}+b)(6-g+v_B) - g\}$$

$$= \max_{0\le b\le \frac{1}{2}} \{-\beta(b)v_A + \beta(b)5 - \gamma(b)\beta(b)g + \beta(b)v_B + \beta(b)\delta(b)\}$$

$$= \max_{0\le b\le \frac{1}{2}} \{\beta(b)\,[3 - \gamma(b)g + \delta(b)]\}$$

For type 2 actions we have the maximization

$$\max_{0\le b\le \frac{1}{2}} \{3\tfrac{1}{2} + (\tfrac{3}{4}-b-b^2)v_A + (\tfrac{1}{8}+b^2)v_{A3} + (\tfrac{1}{8}+b)v_{A4} - v_A - g\}$$

$$= \max_{0\le b\le \frac{1}{2}} \{\beta(b)\,[3 - \gamma(b)g + \delta(b)]\}$$

For the maximization in states B and C we obtain similar equations, which implies that for each state A, B and C we have the same maximization problem, which depends only on the value of g, and of which the outcome is independent of the type of maximizing action chosen. Insertion of the equations for $\beta(b)$, $\gamma(b)$ and $\delta(b)$ in the maximization yields,

$$\max_{0\le b\le \frac{1}{2}} \{3(\tfrac{1}{4}+b+b^2) - g(1\tfrac{1}{4}+b+b^2) + 3\tfrac{5}{8}+b\}$$

$$= \max_{0\le b\le \frac{1}{2}} \{(3-g)b^2 + (4-g)b + (4\tfrac{3}{8}-1\tfrac{1}{4}g)\}.$$

From (9) we see that the expression in braces is a concave parabola in b, hence the maximum is attained at $\tilde{b}(a)$, which is given by

$$\tilde{b}(a) = \frac{4-g}{2(g-3)} = \frac{\frac{5}{8}+a^2}{2(\frac{5}{8}+a)}, \qquad (10)$$

which is in the interval $[0,\frac{1}{2}]$ for all $0\le a\le\frac{1}{2}$.

Any limit point of the PIP (if it exists) has to satisfy the equation $\tilde{b}(a) = a$ and, denoting it by a_0, we have,

$$a_0 = -\frac{5}{8} + \frac{1}{8}\sqrt{65} \approx 0.38$$

If we start with a rational number as the fraction, it follows from equation (10) that at any stage in the policy iteration, the fraction remains rational. Since the limit is a non-rational number, it will take an infinite number of steps to reach it. □

We remark that modifying the policy improvement procedure for the unichain case by replacing equation (2) by

$$v_{i_0}(f^{(n)}) = 0, \qquad n=1,2,\ldots$$

for some state $i_0 \in E$, does not relieve the uniqueness problems. Only if state i_0 is recurrent under all successive policies in the PIP is such an approach possible (cf. [3]). However, such an assumption does not always hold and when it does, it causes only the bias vectors to converge. The second counter example also shows that this is not necessarily the case for the maximizing policies (skipping actions of type 2 in state C causes state A to be recurrent, the maximizing policies still do not converge).

ACKNOWLEDGEMENT

The author would like to thank Arie Hordijk and Martin Puterman for their suggestions and encouragements.

REFERENCES

[1] Puterman, M.L. and S.L. Brumelle. "On the convergence of policy iteration in stationary dynamic programming". Math. Oper. Res. 4, 1979, p.60-69.

[2] Hordijk, A. and M.L. Puterman. "On the convergence of policy iteration in finite state undiscounted Markov decision processes: the unichain case". Math. Oper. Res. 12, 1987, p.

[3] Dekker, R "Denumerable Markov decision chains: optimal policies for small interest rates". Ph.D. thesis. Univ. of Leiden, 1985.

[4] Schweitzer, P.J. "On undiscounted Markovian decision processes with compact action spaces". RAIRO/Operations Research. 19, 1985, p.71-86.

Received: 11/1/1986

Revised: 4/15/1987