The Method of Successive Approximations and Markovian Decision Problems

Arie Hordijk and Henk Tijms

Mathematisch Centrum, Amsterdam, The Netherlands

(Received May 15, 1973)

This note considers HOWARD'S discrete-time Markovian decision model with the average return as criterion. Using results of BLACKWELL AND MACQUEEN for the discounted return model it is shown in all generality that the Odoni bounds contain both the maximal average return and the average return of the current policy.

WE CONSIDER HOWARD'S[4] discrete-time Markovian decision model, where $\{1, \cdots, N\}$ is the set of states, $A(i)$ is the set of actions available in state i, $r(i, a)$ is the immediate return from taking action a while in state i, and $p_{ij}(a)$ is the conditional probability that the system is in state j at time $t+1$, given that the system was in state i at time t and that action a was taken at that time.

The method of successive approximations for solving problems on this model has been studied by several authors.[3,6-11] For the discounted model, MACQUEEN[6] proved that this algorithm supplies bounds on both the maximal expected discounted return and the expected discounted return of the current policy. For the undiscounted model with single-chain aperiodic Markov decision processes, ODONI[8] showed that this algorithm yields bounds on the maximal average return. HASTINGS[3] provided a general proof that the Odoni bounds contain the average return of the current policy. In this note we shall provide a connection between the discounted model and the undiscounted one by showing that, from BLACKWELL'S[1] and MacQueen's[6] results, it can be proved in all generality that the Odoni bounds contain both the maximal average return and the average return of the current policy.

PRELIMINARIES

A STATIONARY POLICY, to be denoted by f, is a policy that prescribes in each state i a single action $f(i) \epsilon A(i)$ whenever the system is in state i. Let F be the class of stationary policies. It is known[1,2] that, for both the total discounted and the average return criteria, we may restrict ourselves to the class F. Let $V_\alpha(i, f)$ be the total expected discounted return starting in state i and using policy f over an infinite horizon, where α with $0<\alpha<1$ is the discount factor. Let $g(i, f)$ be the long-run average expected return per unit time when i is the initial state and policy f is used. Let $V_\alpha(i) = \max_f V_\alpha(i, f)$, and let $g(i) = \max_f g(i, f)$ for $1 \leqq i \leqq N$.

Blackwell[1] proved that there is a policy f^* such that $V_\alpha(i, f^*) = V_\alpha(i)$ for all α close enough to 1 and all i, and $g(i, f^*) = g(i)$ for all i (*see also* pp. 24–25 in DERMAN[2]). From this and part (a) of Theorem 4 in Blackwell,[1] we have, for all i,

$$\lim_{\alpha\to 1}(1-\alpha)V_\alpha(i,f)=g(i,f),\ (f\epsilon F), \text{ and } \lim_{\alpha\to 1}(1-\alpha)V_\alpha(i)=g(i). \quad (1)$$

Let $V_0(i)$, $1\leqq i\leqq N$, be an arbitrary function. For any $\alpha>0$ and $n=0, 1, \cdots$, let

$$V_{n+1}(i,\alpha)=\max_{a\epsilon A(i)}\{r(i,a)+\alpha\textstyle\sum_{j=1}^{j=N} p_{ij}(a)V_n(j,\alpha)\}, \qquad (1\leqq i\leqq N) \quad (2)$$

where $V_0(i,\alpha)=V_0(i)$. For any $\alpha>0$ and $\epsilon\geqq 0$, we define, for $n=0, 1, \cdots$,

$$F_n(\alpha,\epsilon)=\{f\epsilon F|r[i,f(i)]+\alpha\textstyle\sum_{j=1}^{j=N} p_{ij}[f(i)]V_n(j,\alpha)\geqq V_{n+1}(i,\alpha)-\epsilon \text{ for all } i\}.$$

Let $F_n(\alpha)=F_n(\alpha,0)$. For any $0<\alpha<1$ and $\epsilon\geqq 0$, we define, for $n\geqq 0$ and $1\leqq i\leqq N$,

$$u_n'(i,\alpha,\epsilon)=V_n(i,\alpha)+(1-\alpha)^{-1}\min_{1\leqq j\leqq N}\{V_{n+1}(j,\alpha)-V_n(j,\alpha)\}-(1-\alpha)^{-1}\epsilon,$$
$$u_n''(i,\alpha)=V_n(i,\alpha)+(1-\alpha)^{-1}\max_{1\leqq j\leqq N}\{V_{n+1}(j,\alpha)-V_n(j,\alpha)\}.$$

RESULTS

LEMMA 1. *For any $n\geqq 0$ and $1\leqq i\leqq N$, the function $V_n(i,\alpha)$ is continuous in $\alpha>0$.*

Proof. The lemma follows immediately from (2) by induction on n.

LEMMA 2. *Let α^* be a fixed positive number. For any $n\geqq 0$ and $\epsilon>0$, there is a positive number $\alpha_n(\epsilon)$ with $F_n(\alpha^*)\subseteq F_n(\alpha,\epsilon)$ for $|\alpha-\alpha^*|\leqq\alpha_n(\epsilon)$.*

Proof. Fix $n\geqq 0$ and $\epsilon>0$. Assume to the contrary that there is an infinite sequence $\{\alpha_k\}$ such that $\alpha_k\to\alpha^*$ as $k\to\infty$ and, for each k, $F_n(\alpha^*)\backslash F_n(\alpha_k,\epsilon)$ is not empty. Since both F and the set of states are finite, we can now choose a policy f', a state s, and a subsequence $\{\alpha_k'\}$ of $\{\alpha_k\}$ such that

$$r[s,f'(s)]+\alpha_k'\textstyle\sum_j p_{sj}[f'(s)]V_n(j,\alpha_k')<V_{n+1}(s,\alpha_k')-\epsilon \quad \text{for all } k,$$
$$r[s,f'(s)]+\alpha^*\textstyle\sum_j p_{sj}[f'(s)]V_n(j,\alpha^*)=V_{n+1}(s,\alpha^*).$$

Letting $k\to\infty$ and using Lemma 1, we obtain a contradiction.

It is easy to give an example showing that $F_n(\alpha^*)\subseteq F_n(\alpha)$ for all α near α^* need not hold; however, $F_n(\alpha^*)\supseteq F_n(\alpha)$ for all α near α^* does hold.

The next lemma is an immediate extension of results in MacQueen.[6]

LEMMA 3. *For any $0<\alpha<1$, any $\epsilon\geqq 0$ and $n=0, 1, \cdots$,*

$$u_n'(i,\alpha,\epsilon)\leqq V_\alpha(i,f)\leqq V_\alpha(i)\leqq u_n''(i,\alpha) \quad \textit{for all} \quad f\epsilon F_n(\alpha,\epsilon) \textit{ and } 1\leqq i\leqq N, \quad (3)$$

where $u_n'(i,\alpha,\epsilon)$ is nondecreasing in n and $u_n''(i,\alpha)$ is nonincreasing in n.

We are now in a position to prove the desired result.

THEOREM 1. *For any $n=0, 1, \cdots$,*

$$u_n'\leqq g(i,f)\leqq g(i)\leqq u_n'' \quad \textit{for all} \quad f\epsilon F_n(1) \quad \textit{and} \quad 1\leqq i\leqq N,$$

where $u_n'=\min_j\{V_{n+1}(j,1)-V_n(j,1)\}$ and $u_n''=\max_j\{V_{n+1}(j,1)-V_n(j,1)\}$. Also, u_n' is nondecreasing in n and u_n'' is nonincreasing in n.

Proof. Fix $n\geqq 0$ and $\epsilon>0$. Choose β such that $F_n(1)\subseteq F_n(\alpha,\epsilon)$ for $\beta\leqq\alpha<1$ (*see* Lemma 2). Let $f\epsilon F_n(1)$, and let $\alpha\geqq\beta$. Premultiplying each term in (3) by

$1-\alpha$, letting $\alpha \to 1$, and using Lemma 1 and (1), we find $u_n' - \epsilon \leqq g(i, f) \leqq g(i) \leqq u_n''$ for all i. This proves the theorem, since $\epsilon > 0$ was arbitrary.

Theorem 1 may be proved also directly by adapting the proof in Hastings.[3]

Remark. Consider the case where, for every average return optimal policy f, the Markov matrix $\{p_{ij}[f(i)]\}$ is aperiodic and single-chained. Then, $g(i)$ is constant (say g) and $V_n(i, 1) - ng$ has a finite limit for all i as $n \to \infty$.[5,9] From this it is readily verified that u_n' and u_n'' both converge to g as $n \to \infty$ and that there is an integer n_0 such that, for all $n \geqq n_0$, every policy $f \epsilon F_n(1)$ is average return optimal (cf. Odoni[8]).

REFERENCES

1. D. Blackwell, "Discrete Dynamic Programming," *Ann. Math. Stat.* **33,** 719–726 (1962).
2. C. Derman, *Finite State Markovian Decision Processes,* Academic Press, New York, N.Y., 1970.
3. N. A. J. Hastings, "Bounds on the Gain of a Markovian Decision Process," *Opns. Res.* **19,** 240–243 (1971).
4. R. A. Howard, *Dynamic Programming and Markov Processes,* The M.I.T. Press, Cambridge, Mass., 1960.
5. E. Lanery, "Étude Asymptotic des Systèmes Markoviens à Commande," *Revue Inf. Rech. Operat.* **1,** No. 5, 3–57 (1967).
6. J. B. MacQueen, "A Modified Dynamic Programming Method for Markovian Decision Problems," *J. Math. Anal. and Appl.* **14,** 38–43 (1966).
7. T. E. Morton, "Undiscounted Markov Renewal Programming via Modified Successive Approximations," *Opns. Res.* **19,** 1081–1089 (1971).
8. A. R. Odoni, "On Finding the Maximal Gain for Markov Decision Processes," *Opns. Res.* **17,** 857–860 (1969).
9. P. J. Schweitzer, "Perturbation Theory and Markovian Decision Processes," M.I.T. Operations Research Center Technical Report No. 15, 1965.
10. ———, "Multiple Policy Improvements in Undiscounted Markov Renewal Programming," *Opns. Res.* **19,** 784–793 (1971).
11. D. J. White, "Dynamic Programming, Markov Chains, and the Method of Successive Approximations," *J. Math. Anal. and Appl.* **6,** 373–376 (1963).