Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Technical Note—Bounds on the Gain of a Markov Decision Process

N. A. J. Hastings,

minimizes the total expected cost. For the test sequence $(j_1, \cdots, j_n)$, this expected cost is given by

$$F(j_1, \cdots, j_n) = C_{j_1} + C_{j_2}(1-R_{j_1}) + C_{j_3}(1-R_{j_1})(1-R_{j_2}) + \cdots + C_{j_n}\prod_{k=1}^{n-1}(1-R_{j_k})$$
$$= \sum_{m=1}^{m=n} C_{j_m} g_m(j_1, \cdots, j_m),$$

where we define $g_1(j_1) = 1$ and $g_i(j_1, \cdots, j_i) = \prod_{k=1}^{i-1}(1-R_{j_k})$ for $1 < i \leqq n$. Conditions (A) and (B) are easily verified and, in particular, we have $G(y_j) = R_{y_j}$ and $G_i(x_1, \cdots, x_{i-1}) = \prod_{k=x_1}^{k=x_{i-1}}(1-R_k)$. Therefore, the least-cost testing sequence is provided by performing the tests in the order $j_1^*, j_2^*, \cdots, j_n^*$, where $C_{j_1^*}/R_{j_1^*} \leqq C_{j_2^*}/R_{j_2^*} \leqq \cdots \leqq C_{j_n^*}/R_{j_n^*}$.

REFERENCES

1. R. Bellman, *Dynamic Programming*, Princeton University Press, 1957.
2. D. C. Denby, "Minimum Downtime as a Function of Reliability and Priority Assignments in a Component Repair," *J. Indust. Eng.* **18,** 436–439 (1967).
3. H. Greenberg, "Optimum Test Procedure Under Stress," *Opns. Res.* **12,** 689–692 (1964).
4. B. Gluss, "An Optimum Policy for Detecting a Fault in a Complex System," *Opns. Res.* **7,** 468–477 (1959).
5. L. G. Mitten, "An Analytic Solution to the Least Cost Testing Sequence Problem," *J. Indust. Eng.* **11,** 17 (1960).
6. W. E. Smith, "Various Optimizers for Single-Stage Production," *Nav. Res. Log. Quart.* **3,** 59–66 (1956).

BOUNDS ON THE GAIN OF A MARKOV DECISION PROCESS

N. A. J. Hastings

University of Birmingham, Birmingham, England

(Received January 8, 1970)

An algorithm for the steady-state solution of Markov decision problems has been proposed by Howard and modified by Hastings. This note shows, for the case of single-chain Markov decision processes, how bounds on the optimal gain can be obtained at each cycle of the foregoing algorithms. The results extend to Markov renewal programming. Related results are the bounds proposed by Odoni for use with White's value-iteration method of optimization.

An algorithm for the steady-state solution of Markov decision problems has been proposed by Howard[4] and modified by Hastings.[2, 3] This note shows, for the case of single-chain processes, how bounds on the optimal gain can be obtained at each cycle of the foregoing algorithms. The results extend to Markov renewal programming. Related results are the bounds proposed by ODONI[6] for use with the value-iteration method of optimization described by BELLMAN[1] and WHITE.[7]

DISCRETE PROCESSES

THE PROPOSED bounds and their derivation are most readily developed for Howard's algorithm and a discrete Markov process. Extensions to Hastings's algorithm and to continuous (Markov-renewal) processes then follow.

Consider a system that has a finite number of discrete states, $i=1, 2, \cdots, N$, and undergoes a completely ergodic, single-chain Markov decision process in discrete time. At each state, a number of alternative actions k are available. Policy A (vector) consists of a particular set of actions one for each state. Application of policy A gives the system a steady-state gain g^A per stage and gives each state a relative value v_i^A. Under action k at state i there is a mean immediate return q_i^k and the probability of transition from state i to state j is p_{ij}^k.

Both algorithms consist of repeated cycles of value-determination operation and policy-improvement routine. In the value-determination operation, the gain and state values under a current policy A are found. In the policy-improvement routine, an improved policy B is found. For Howard's algorithm, improved policies are found by maximizing the test quantity

$$\max_k \left[q_i^k+\sum_{j=1}^{j=N} p_{ij}^k v_j^A\right]. \tag{1}$$

For any two policies X, Y, let

$$y_i^{XY}=q_i^Y-q_i^X+\sum_{j=1}^{j=N} p_{ij}^Y v_j^X-\sum_{j=1}^{j=N} p_{ij}^X v_j^X.$$

Then y_i^{AB} is the improvement in the test quantity at stage i during the policy-improvement routine. Let policy M be an optimal policy, and let Π_i^M be the steady-state probability for state i under policy M.

The bounds on the optimal gain are

$$g^A+\min_i[y_i^{AB}]\leqq g^B\leqq g^M\leqq g^A+\max_i[y_i^{AB}]. \tag{2}$$

The derivation of this result is as follows. Following Howard (reference 4, page 42) we can show that, for any two policies, and in particular for an optimal policy M and the current policy A,

$$g^M-g^A=\sum_{i=1}^{i=N} \Pi_i^M y_i^{AM}. \tag{3}$$

Thus, the difference in the gains is equal to a weighted average of the y_i^{AM} terms and is not greater than the largest, $g^M-g^A\leqq\max_i[y_i^{AM}]$. But from (1) for all i, $y_i^{AB}\geqq y_i^{AM}$, and hence the upper bound follows.

The optimal gain g^M is greater than or equal to g^B by definition. The superiority of g^B to the left-hand expression in (2) is established from the equation

$$g^B-g^A=\sum_{i=1}^{i=N} \Pi_i^B y_i^{AB}.$$

Hastings's algorithm is similar to Howard's, except that, during the policy-improvement routine, improved state values are introduced. Given a current policy

A, an improved policy B is found using the test quantity shown on the right-hand side of equation (4). This equation also defines the improved value v_i^{AB}:

$$v_i^{AB}+g^A=\max_k\,[q_i{}^k+\textstyle\sum_{j=1}^{i-1} p_{ij}^k v_j^{AB}+\sum_{j=i}^{j=N} p_{ij}^k v_j{}^A]. \tag{4}$$

For any two policies X, Y, define v_i^{XY} by

$$v_i^{XY}+g^X=q_i{}^Y+\textstyle\sum_{j=1}^{i-1} p_{ij}^Y v_j^{XY}+\sum_{j=i}^{j=N} p_{ij}^Y v_j{}^X,$$

and let $z_i^{XY}=v_i^{XY}-v_i{}^X$. The bounds are

$$g^A+\min_i[\textstyle\sum_{j=i}^{j=N} p_{ij}^B z_j^{AB}]\leqq g^B\leqq g^M\leqq g^A+\max_{i,\,X}\,[\sum_{j=i}^{j=N} p_{ij}^X z_j^{AX}]\leqq g^A+\max_i[z_i^{AB}].$$

The derivation is similar to that for Howard's algorithm. The result corresponding to (4) is given by Hastings (reference 2, equation 18), and is, in the current notation,

$$g^M-g^A=\textstyle\sum_{i=1}^{i=N}\Pi_i\sum_{j=i}^{j=N} p_{ij}^M z_j^{AM}.$$

Since the policy-improvement routine maximizes z_j and not $\sum_{j=i}^{j=N} p_{ij}z_j$, we have the relation $z_j^{AB}\geqq z_j^{AM}$, but not $\sum_{j=i}^{j=N} p_{ij}^B z_j^{AB}\geqq\sum_{j=i}^{j=N} p_{ij}^M z_j^{AM}$, and so the lesser upper bound includes a maximization over all policies X.

MARKOV RENEWAL PROGRAMMING

The extension of discrete Markov decision problems to the continuous case is described by Jewell.[5]

Consider a process similar to that already described but occurring in continuous time. Let g^A denote the gain per unit time for policy A; let $T_i{}^k$, $T_i{}^A$ denote the mean duration of visits to state i under action k and policy A, respectively; let $\Pi_i{}^A$ denote the steady-state probability that the system is in state i. The interpretations of the symbols for the stage returns, transition probabilities, and state values are the same as before.

Policy optimization can be achieved by Howard's algorithm using either of two types of test quantities, and by Hastings's algorithm using a modification of one of these test quantities. For Howard's algorithm, the test quantities, given a current policy A, are,

$$\max_k\,[(q_i{}^k+\textstyle\sum_{j=1}^{j=N} p_{ij}^k v_j{}^A-v_i{}^A)/T_i{}^k], \tag{5}$$

$$\max_k\,[q_i{}^k-g^AT_i{}^k+\textstyle\sum_{j=1}^{j=N} p_{ij}^k v_j{}^A]. \tag{6}$$

For any two policies X, Y, let

$$y_i^{XY}=q_i{}^Y-g^XT_i{}^Y+\textstyle\sum_{j=1}^{j=N} p_{ij}^Y v_j{}^X-v_i{}^X.$$

If test quantity (5) is used, the bounds are

$$g^A+\min_i\,[y_i^{AB}/T_i{}^B]\leqq g^B\leqq g^M\leqq g^A+\max_i\,[y_i^{AB}/T_i{}^B]. \tag{7}$$

The derivation of this result is exactly parallel to the derivation of (2) and starts from the continuous-time equation corresponding to (3), which is

$$g^M-g^A=\textstyle\sum_{i=1}^{i=N}\Pi_i{}^M(y_i^{AM}/T_i{}^M).$$

If the test quantity (6) is used, the lower bounds are unchanged from (7), but the upper bound is weaker; it is

$$g^M \leqq g^A + \max_{i,X} [y_i^{AX}/T_i^X]. \tag{8}$$

Maximization over all policies X must be retained in (8) because the policy-improvement routine now maximizes y_i and not y_i/T_i.

Hastings's algorithm has been developed only for a modification of the second test quantity (6) to the form,

$$v_i^{AB} = \max_k [q_i^k - g^A T_i^k + \sum_{j=1}^{i-1} p_{ij}^k v_j^{AB} + \sum_{j=i}^{j=N} p_{ij}^k v_j^A].$$

For any two policies X, Y, let

$$v_i^{XY} = q_i^Y - g^X T_i^Y + \sum_{j=1}^{i-1} p_{ij}^Y v_j^{XY} + \sum_{j=i}^{j=N} p_{ij}^Y v_j^X, \; z_i^{XY} = v_i^{XY} - v_i^X.$$

The bounds on g^M are

$$g^A + \min_i [\sum_{j=i}^{j=N} p_{ij}^B z_j^{AB}/T_j^B] \leqq g^B \leqq g^M \leqq g^A + \max_{i,X} [\sum_{j=i}^{j=N} p_{ij}^X z_j^{AX}/T_j^X].$$

The derivation of these bounds uses similar arguments to those already developed for other cases.

With regard to Odoni's valuable paper, some further points are noteworthy.

First, it is implicit, but should perhaps be stated, that, in the value-iteration method, the steady-state gain g^A under the current policy A found at stage n is bounded by Odoni's bounds:

$$\min_i [x_i(n)] \leqq g^A \leqq \max_i [x_i(n)].$$

This follows by inserting $v(n+1) = v(n) + x(n)$ into Odoni's equation $v(n+1) = q^A + p^A v(n)$ and taking scalar products with Π^A. Hence, if the bounds are close, the current policy can be used as a good suboptimal policy.

Second, the value-iteration method may actually converge, and convergence occurs when all the $x_i(n)$ are equal. The speed of convergence is increased if the boundary values $v_i(0)$ are chosen to be as close as possible to the steady-state relative values v_i^M under an optimal policy. The v_i^M are, of course, unknown, but it is not unusual in practice to find that the optimal relative values of the states can be guessed to a reasonable degree of accuracy.

ACKNOWLEDGMENT

THE EXTENSION of these bounds to Markov renewal programming was suggested to me by PAUL J. SCHWEITZER, who also derived the other bounds independently.

REFERENCES

1. R. E. BELLMAN, "A Markovian Decision Process," *J. Math. Mech.* **6,** 679–684 (1957).
2. N. A. J. HASTINGS, "Some Notes on Dynamic Programming and Replacement," *Opnl. Res. Quart.* **19,** 453–464 (1968).
3. ———, "Optimisation of Discounted Markov Decision Problems," *Opnl. Res. Quart.* **20,** 499–500 (1969).
4. R. A. HOWARD, *Dynamic Programming and Markov Processes,* Wiley, New York, 1960.

5. W. S. JEWELL, "Markov-Renewal Programming, I and II," *Opns. Res.* **11**, 938–971 (1963).
6. A. R. ODONI, "On Finding the Maximal Gain for Markov Decision Processes," *Opns. Res.* **17**, 857–860 (1969).
7. D. J. WHITE, "Dynamic Programming, Markov Chains and the Method of Successive Approximations," *J. Math. Anal. and Appl.* **6**, 373–376 (1963).

ON THE ASYMPTOTIC CONVERGENCE RATE OF COST DIFFERENCES FOR MARKOVIAN DECISION PROCESSES

Thomas E. Morton

Carnegie-Mellon University, Pittsburgh, Pennsylvania

(Received January 2, 1970)

The modified method of successive approximations for solving Markovian decision problems as formulated by WHITE, SCHWEITZER, MACQUEEN, and ODONI, concentrates attention on cost differences either between successive stages in the same state, or relative to a base state in the same stage, rather than on the total cost function. The former bound the (discounted) gain of the optimal policy, while the latter relative-cost function determines the policy to be chosen at each stage. While these authors have demonstrated that these modified constructs converge to the gain and the optimal relative-cost function under rather general circumstances (undiscounted, single-chain, aperiodic processes), little is known about the rates of convergence. [Note that convergence of the relative-cost function guarantees optimality of a currently repeating policy, as noted by HOWARD.] A great deal of insight into this mathematically difficult question may be gained by working out the actual asymptotic convergence rates of these constructs for the special case of a single fixed policy. This is an easy exercise via Howard's methods, but very suggestive, since the policy *will* be asymptotically constant for a well-behaved problem. (In particular, if there is a unique optimal policy it will eventually repeat.) Convergence for both constructs for the fixed-policy case is very powerful even for discount rates greater than 1.0, depending principally on the dominant eigenvalue of the transition matrix. This note discusses the intuitive implications of this fact for the relative efficiencies of modified value iteration, policy iteration, policy iteration via successive approximations, or possible hybrids.

WHITE[6] FIRST modified the method of successive approximations for solving Markov decision problems to focus attention on the convergence of costs relative to the cost of a base state rather than on convergence of the total cost function. For the undiscounted case, he proved rather elegantly that the modified cost function converged at least geometrically $\sim(1-\gamma)^{n/k}$, where n is the iteration, and there is postulated to be a state that one must return to every k iterations with probability at least γ, irrespective of the sequence of policies chosen. He realized that the true convergence rate might be much faster, and that the latter

Published in Operations Research on February 01, 1971 as DOI: 10.1287/opre.19.1.240.
This article has not been copyedited or formatted. The final version may differ from this version.