Operations Research

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

Technical Note—On the Asymptotic Convergence Rate of Cost Differences for Markovian Decision Processes

Thomas E. Morton,

To cite this article:
Thomas E. Morton, (1971) Technical Note—On the Asymptotic Convergence Rate of Cost Differences for Markovian Decision Processes. Operations Research 19(1):244-248. https://doi.org/10.1287/opre.19.1.244

Published in Operations Research on February 01, 1971 as DOI: 10.1287/opre.19.1.244.
This article has not been copyedited or formatted. The final version may differ from this version.

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.19.1.244.
This article has not been copyedited or formatted. The final version may differ from this version.

restriction might be relaxed. SCHWEITZER[5] proved convergence for the more general single-chain aperiodic case, but said very little about the rate of convergence.

MACQUEEN[2] and ODONI[4] have greatly strengthened the position of the user of the procedure by modifying White's method slightly to provide computable upper and lower bounds on the gain of the process at each iteration.

There is one special case for which it is easy to get sharp estimates of the asymptotic behavior of these modified costs and bounds via HOWARD's methods[1]—the special case of a single fixed policy. For *arbitrary* discount factor, the gain bounds for a fixed policy turn out to converge asymptotically geometrically with a factor of β, where β is the dominant eigenvalue of the transition matrix. Thus convergence of these bounds is independent of the discount rate. Similarly the relative cost function for this special case converges at rate $\alpha\beta$, where α is the discount factor. Thus the relative cost function converges for discount factors greater than or equal to 1.0 as long as $\alpha < 1/\beta$. Of course, in practice one need not actually compute β in advance to use this fact; attainment of geometric convergence is easily recognizable as it occurs. These facts will be demonstrated in the next section (utilizing Odoni's definitions and notation as much as possible).

While these results are not particularly deep, it does not seem that the following practical and intuitive implications have been drawn.

First, if there is a unique optimal policy, experience of many authors indicates that it is often reached very early. One is indeed, after that point, iterating a fixed policy. It is necessary for the relative costs to converge to guarantee that this policy will in fact be optimal, so that the fixed-policy asymptotic convergence is of great interest.

Second, if the problem is 'well-behaved' in the sense that policies with gains close to that of the optimal policy also have transition matrices very similar to that of the optimal policy, then one would expect convergence characteristics 'close to' that for a fixed policy. Since policies with gains not close to optimal should be eliminated fairly early, it is at least quite plausible, then, that the over-all number of iterations be governed by the dominant eigenvalue also. (The author predicts, for example, that many problems will converge by modified successive approximations for discount rates greater than 1.0.)

Third, these results make it absolutely clear that, even for the policy iteration method, it is inefficient both from the point of view of computer storage, and computations required to do the 'policy-evaluation' step by solving a set of simultaneous equations. Simply repeatedly using the successive approximation machinery with the policy kept fixed will produce both the gain and the needed relative cost function with geometric convergence. (Note that this machinery is needed for the 'policy-improvement' step in any event, so that the programming needed for such a 'modified policy iteration' is actually a subset of that for ordinary policy iteration.)

Finally, mixtures of the two procedures may be employed. For example, one could alternate a full maximizing iteration with 3 or 4 'cheap' fixed iterations to approximate the relative cost function of the current policy to increase the chance that the next maximizing iteration would find an improved policy, etc. These ideas will be discussed more fully in the final section.

Published in Operations Research on February 01, 1971 as DOI: 10.1287/opre.19.1.244.
This article has not been copyedited or formatted. The final version may differ from this version.

ASYMPTOTIC CONVERGENCE FOR A FIXED POLICY

A FINITE-STATE, discrete-time Markov system is controlled by a decision-maker. After each transition $n=0, 1, 2, \cdots$, the system is in one of N states $i=1, \cdots, N$. Then the decision-maker picks one of K actions resulting in an expected reward for the period of $q_i^{(k)}$ and the transition probability row vector $P_i^{(k)}=[p_{ij}^{(k)}]$, $j=1, \cdots, N$, where $[p_{ij}^{(k)}]$ is the probability of going from state i to state j given decision k. We define $v_i(n)$ as the total expected earnings from the next n transitions, if the system is now in state i, and if an optimal policy is followed. Future costs are discounted by the factor α, where the only initial restriction on α is $-\infty<\alpha<\infty$.

Then we can write:

$$v_i(n+1)=\max_k \{q_i^{(k)}+\alpha P_i^{(k)} v(n)\}. \tag{1}$$

Now we restrict our attention to the special case of a single possible policy, for which this easily simplifies (after Howard) to

$$v(n+1)=\sum_{i=0}^{i=n} (\alpha P)^i q+(\alpha P)^{n+1} v(0). \tag{2}$$

The inclusion of the 'salvage' term $(\alpha P)^{n+1}v(0)$ allows us to interpret (2) as the terminal stage of a full dynamic programming problem $(n+1)$ iterations after the optimal policy (with transition matrix P and one stage reward vector q) has been permanently achieved. (It should be repeated that nothing is being proved here about the achievement of such a position in the first place.)

Suppose for ease of exposition that P has distinct real eigenvalues $1, \beta_1, \cdots, \beta_{N-1}$ where $1>|\beta_1|>|\beta_2|\cdots$, etc. (It is not really necessary that the eigenvalues be real or distinct, only that $|\beta_i|<1$ and that there be only one ergodic class. However, these generalizations would really add little and might tend to obscure the point.)

Then after Howard (reference 1, pages 9–12).

$$P^i=S+\sum_{j=1}^{N-1} \beta_j{}^i T_j, \tag{3}$$

where S has identical rows that are the stationary probabilities and the T_i are transient matrices. It is convenient to define

$$S_q=k\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}=ke, \quad (T_j)q=q_j, \quad Sv(0)=ke, \quad (T_j)v(0)=r_j. \tag{4}$$

Thus k is the generalized gain, while the q_j, r_j are transient-cost error vectors. Thus equation (2) can be rewritten as

$$v(n+1)=\{\sum_{i=1}^{i=n} \alpha^i(ke+\sum_{j=1}^{N-1} \beta_j{}^i q_j)\}+\alpha^{n+1}k_0 e+\alpha^{n+1}\sum_{j=1}^{N-1} \beta_j^{n+1} r_j. \tag{5}$$

Thus if, in the fashion of Odoni we define

$$x(n)=\alpha^{-n}[v(n+1)-v(n)], \tag{6}$$

it follows directly that

$$x(n)=[k+(\alpha-1)k]e+\beta_1{}^n[q_1+(\alpha\beta_1-1)r_1]+O(\beta_2{}^n). \tag{7}$$

Now let $\bar{q}$ be the largest component of $[q_1+(\alpha\beta_1-1)r_1]$ and $\underline{q}$ be the smallest, and

let $K \equiv k+(\alpha-1)k_0$. Then the bounds of Odoni for the generalized gain become

$$\begin{aligned} L''(n) &\equiv [\max_i x_i(n)] \sim K+\beta_1{}^n\bar{q}, \\ L'(n) &\equiv [\min_i x_i(n)] \sim K+\beta_1{}^n\bar{q}. \end{aligned} \tag{8}$$

Thus, the number of iterations required for convergence of x, L', or L'' is of the order of $1/(1-\beta_1)$ irrespective of α.

Next we turn to the convergence of the relative cost function defined (again as by Odoni):

$$w(n) = v(n) - v_N(n)e. \tag{9}$$

Now, if we define

$$q_i' \equiv q_i - q_N, e, \qquad r_i' = r_i - r_N, e \tag{10}$$

(that is, subtract the last component from every component), we have from (5) and (9) that

$$\begin{aligned} w(n) &= \textstyle\sum_{i=0}^{n-1} \alpha^i \{(ke+\sum_{j=1}^{N-1} \beta_j{}^i q_j) - (k+\sum_{j=1}^{N-1} \beta_j{}^i q_{Nj})e\} + \sum_{j=1}^{N-1} (\alpha\beta_j)^n r_j' \\ &= \textstyle\sum_{j=1}^{N-1} \{[1-(\alpha\beta_j)^n/(1-\alpha\beta_j)]q_j' + (\alpha\beta_j)^n r_j'\}. \end{aligned} \tag{11}$$

Thus, the limiting relative-cost vector is given by

$$\bar{w} = w(\infty) = \textstyle\sum_{j=1}^{N-1} q_j'(1-\alpha\beta_j)^{-1}, \tag{12}$$

and the principal transient is given by

$$q = -q_1'(1-\alpha\beta_1)^{-1} + r_1', \tag{13}$$

allowing us to simplify the n-stage relative cost function to

$$w(n) = \bar{w} + (\alpha\beta_1)^n\tilde{q} + O[(\alpha\beta_2)^n]. \tag{14}$$

Thus, the convergence of the relative-cost function takes about $1/(1-\alpha\beta_1)$ iterations, where we must restrict

$$-(1/\beta_1) < \alpha < (1/\beta_1). \tag{15}$$

However, α can always be 1.0 and usually considerably larger, and relative costs will still converge.

COMPUTATIONAL IMPLICATIONS

For a typical large-scale problem,[3] one does not need to calculate and store all possible transition matrices and costs. Instead, one stores functional relations for calculating the one-stage cost as a function of the state and the decision, which states can be reached from the current state, and the associated probabilities. These require very little storage, so that needed storage is three or four times the number of states N. If from each state there are say at most J possible transitions and K possible decisions, it is almost always the case that both J and K are much smaller than N.

Ordinary value iteration takes on the order of $1/(1-\alpha)$ iterations for the cost function to converge, so that the total computational effort is on the order of

Published in Operations Research on February 01, 1971 as DOI: 10.1287/opre.19.1.244.
This article has not been copyedited or formatted. The final version may differ from this version.

$JKN/(1-\alpha)$. In particular, of course, convergence is tortuous for α close to 1.0 and impossible for $\alpha = 1.0$.

The method of checking cost differences, which we might call 'modified value iteration,' would take, after convergence to the optimal policy, about $JKN/(1-\alpha\beta^*)$ computations. From the author's experience, initial convergence to the optimal policy requires a similar amount of effort. Problems of similar structure seem to have similar dominant eigenvalues regardless of N.

The value-determination step of policy iteration requires N^3 calculations per iteration and vastly more storage than successive approximations. No amount of decomposition or fancy programming would seem able to make large problems tractable by this method.

On the other hand, policy iteration features the comfort of guaranteed monotone convergence in a finite number of policy guesses. If the value-determination step were to be done by modified successive approximations, the value determination step would be reduced to $JN/(1-\alpha\beta^*)$ computations, and storage requirements down to $3N$ or $4N$. The modified policy iteration and modified value iteration would both still require JKN calculations on the full maximizing step. For large values of K it might very well be that $JN/(1-\alpha\beta^*) \ll JKN$, so that the value determination step might become relatively cheap. Then one would be able to compare modified value iteration and modified policy iteration directly by the number of major iterations required.

Taking this one step further, a reasonable hybrid scheme might be: one full iteration alternating with 5 or 6 'cheap' fixed iterations in the early stages when straight modified value iteration might converge slowly, one full iteration alternating with one or two in the middle stages, switching completely to the cheap iterations after the same policy began to repeat, terminated by a full iteration just to check the policy.

REFERENCES

1. R. A. Howard, *Dynamic Programming and Markov Processes,* The MIT Press, Cambridge, 1960.
2. J. B. MacQueen, "A Modified Dynamic Programming Method for Markovian Decision Problems," *J. Math. Anal. and Appl.* **14,** 38–43 (1966).
3. T. E. Morton, "The Near Myopic Nature of the Lagged Proportional Cost Inventory Problem with Lost Sales," Carnegie-Mellon University, Graduate School of Industrial Administration Working Paper, W. P. 19-69-7.
4. A. R. Odoni, "On Finding the Maximal Gain for Markov Decision Processes," *Opns. Res.* **17,** 857–860 (1969).
5. P. J. Schweitzer, "Perturbation Theory and Markovian Decision Processes," MIT Operations Research Center Technical Report, No. 15, June, 1965.
6. D. J. White, "Dynamic Programming Markov Chains and the Method of Successive Approximations," *J. Math Anal. and App.* **6,** 373–376 (1963).

Published in Operations Research on February 01, 1971 as DOI: 10.1287/opre.19.1.244.
This article has not been copyedited or formatted. The final version may differ from this version.