Iterative Solution of the Functional Equations of Undiscounted Markov Renewal Programming

PAUL J. SCHWEITZER

Institute for Defense Analyses, Arlington, Virginia 22202

Submitted by Richard Bellman

An iterative procedure is described for finding a solution of the functional equations

\[ v_i^* = \max_k \left[ q_i^k - g^* T_i^k + \sum_{j=1}^{N} P_{ij}^k \psi_j^* \right] \quad 1 \leq i \leq N \]

of undiscounted Markov renewal programming. It generalizes the method of successive approximations proposed by D. J. White for Markovian decision processes, and requires additional care to ensure stability.

1. INTRODUCTION

This paper provides an iterative solution procedure for the functional equations of Markov renewal programming. These equations are

\[ v_i^* = \max_{k \in \alpha_i} \left[ q_i^k - g^* T_i^k + \sum_{j=1}^{N} P_{ij}^k \psi_j^* \right] \quad 1 \leq i \leq N \]  \hspace{1cm} (1)

or, equivalently,

\[ g^* = \max_{k \in \alpha_i} \left[ q_i^k + \sum_{j=1}^{N} P_{ij}^k \psi_j^* - v_i^* \right] / T_i^k \quad 1 \leq i \leq N, \]

where

\[ T_i^k > 0, \quad P_{ij}^k \geq 0, \quad \text{and} \quad \sum_{j=1}^{N} P_{ij}^k = 1. \]

They arise in \( N \)-state Markov renewal processes with undiscounted rewards [1-4]. Here \( \alpha_i \) denotes the prescribed finite set of alternatives in state \( i \), while \( q_i^k, T_i^k, \) and \( P_{ij}^k \) denote the prescribed mean one-transition reward, mean holding time, and transition probability to state \( j \), if alternative \( k \) is employed in state \( i \).
It is assumed that, for each policy $A e X$, the associated stochastic matrix $P^A$ has one subchain (closed, irreducible set of states) plus possibly some transient states feeding this subchain. With this assumption, Eq. (1) always has a solution [1–4], with $g^*$ unique and $v_i^*$ unique up to one additive constant [5, Corollary to Theorem 2].

These equations are of interest because $g^*$ is the maximal gain rate and because any policy achieving all $N$ maxima in Eq. (1) attains the maximum gain rate [5, Corollary to Theorem 2]. Thus solution of Eq. (1) permits determination of optimal policies.

Published optimization techniques for undiscounted Markov renewal processes fall into two classes, dynamic and linear programming. Dynamic programming approaches [1–4] employ approximation in policy space to solve Eq. (1), and require repeated solutions of sets of $N$ simultaneous equations. Linear programming approaches [6] and [7, Primal III] for determining $g^*$ are likewise impractical if $N$ is large.

A third approach, solution of (1) by the method of successive approximations, is presented here. It appears practical and competitive (Ref. [8]) for extremely large $N$ (thousands) and can be easily computer-coded without the severe storage problems arising in the $DP$ or $LP$ approaches. The iterative scheme generalizes the one used by White [9] for Markov decision processes, with special care taken to avoid instability.

2. White's Iterative Scheme

In the special case where $T_i^k = 1$ for all $i$ and $k$ (Markov case), White [9] has proposed a rapidly converging iterative scheme for (1–2). He sets $v_{N^*} = 0$ (without loss of generality), rewrites (1–2) as

$$g^* = \max_{k \in \alpha N} \left[ q_N^k + \sum_{j=1}^{N-1} P_{kj} v_j^* \right]$$

$$v_i^* = \max_{k \in \alpha i} \left[ q_i^k - g^* + \sum_{j=1}^{N-1} P_{ij} v_j^* \right], \quad 1 \leq i \leq N - 1$$

and solves (3) by successive approximations:

$$g(n) = \max_{k \in \alpha - N} \left[ q_N^k + \sum_{j=1}^{N-1} P_{kj} v(n)_j \right]$$

$$v(n + 1)_i = \max_{k \in \alpha i} \left[ q_i^k - g(n) + \sum_{j=1}^{N-1} P_{ij} v(n)_j \right], \quad 1 \leq i \leq N - 1$$
If a fairly stringent condition on all the multistep transition probabilities is met, White showed that Eq. (4) converges, with \( g(n) \rightarrow g^* \) and \( v(n)_i \rightarrow v_i^* - v_N^* = v_i^* \). Schweitzer [4, Theorem 10.5] established this result under the weaker hypothesis that, for every maximal-gain policy \( A \) (or for every policy \( A \)), \( P^A \) has a single subchain and all its recurrent states are aperiodic.

3. Generalization of White’s Scheme

How can White’s scheme be modified for the general case of unequal \( T_i^k \)? Straightforward generalization of Eq. (4) will not suffice, because the iterative scheme

\[
g(n) = \max_{k \in \mathbb{N}} \left[ q_N^k + \sum_{j=1}^{N-1} P_{Nj}^k v(n)_j \right] / T_N^k \quad (5a)
\]

\[
v(n+1)_i = \max_{k \in \mathbb{N}} \left[ q_i^k - g(n) T_i^k + \sum_{j=1}^{N-1} P_{ij}^k v(n)_j \right] \quad 1 \leq i \leq N - 1 \quad (5b)
\]

is unstable. An example of nonconvergence is given by the 2-state case with one alternative per state. These equations take the form \( v(n+1)_1 = r + sv(n)_1 \) where \( r = q_1 - q_2 T_1/T_2 \) and \( s = P_{11} - P_{21} T_1/T_2 \). If \( T_2/P_{21} \) is sufficiently small, \( s < -1 \) and \( v(n+1)_1 \) diverges with ever-growing oscillations. Our remedy will be to “sample” the process at time intervals \( \tau \) shorter than \( T_2/P_{21} \) or \( T_1/P_{12} \).

4. Proposed Iterative Scheme

A stable modification of White’s scheme, to be motivated below, is the following:

\[
g(n) = \max_{k \in \mathbb{N}} \left[ q_N^k + \sum_{j=1}^{N-1} P_{Nj}^k v(n)_j \right] / T_N^k \quad (6a)
\]

\[
v(n+1)_i = v(n)_i + \tau \max_{k \in \mathbb{N}} \left[ q_i^k + \sum_{j=1}^{N-1} P_{ij}^k v(n)_j - v(n)_i - g(n) T_i^k \right] / T_i^k \quad 1 \leq i \leq N - 1 \quad (6b)
\]
where $\tau$ is a positive constant, with dimensions of time, chosen sufficiently small that

$$(1 - P_{ii}^k) \frac{\tau}{T_i^k} < 1 \quad \text{all } i, k. \quad (7)$$

The proof given below shows that this iterative scheme will always converge if $\tau$ satisfies (7) and if, for every maximal-gain policy $A$ (or, for every policy $A$), $P^A$ has a unique subchain. (Note that $P^A$ is permitted to have periodic and transient states.) Comparison with (1-2) shows that the limits are unique, namely $g(\infty) = g^*$ and $\psi(\infty)_i = v_i^* - v_N^* = v_i^*, \quad 1 \leq i \leq N - 1.$

In the Markov case, where $T_i^k = 1$, convergence of Eq. (6) will occur for any $\tau < 1$. (In general, Eq. (7) is satisfied whenever $\tau < \min_{i,k} T_i^k$). The choice $\tau = 1$ recovers Eq. (4) from Eq. (6), and convergence will occur if every $P^A$ is both unichained and aperiodic. However, if periodic states exist, this scheme with $\tau = 1$ (and White’s) need not converge. (For example, if $N = 2$, $q_i^k = 0$ and $P_{ij}^k = 1 - \delta_{ij}$, then convergence will not occur for arbitrary $\psi(0)$.) If uncertainty exists about the existence of periodic states, Eq. (6) provides a preferable scheme to White’s for the Markov case, because convergence is guaranteed for any $\tau < 1$.

Convergence to maximal-gain policies is also achieved by Eq. (6). Since $g(n)$ and $\psi(n)$ converge to $g^*$ and $\psi^*$, any policy achieving the $N$ maxima in Eq. (6) for infinitely many $n$ must achieve the $N$ maxima in Eq. (1), hence be maximal-gain. Consequently, for any $n > n_0$ (where the first $n_0$ iterations of Eq. (6) contain all of the maximizing policies which appear finitely often) any policy achieving the $N$ maxima in Eq. (6) is maximal-gain.

5. PROOF OF CONVERGENCE

Convergence of this scheme is demonstrated by rewriting it in a form paralleling White’s. Define fictitious transition probabilities, indicated by a tilde, by

$$\tilde{P}_{ij}^k = \delta_{ij} + (P_{ij}^k - \delta_{ij}) \frac{\tau}{T_i^k}.$$ 

Then

$$\tilde{P}_{ij}^k = \frac{P_{ij}^k}{T_i^k} \geq 0 \quad i \neq j, \quad (8a)$$

$$\tilde{P}_{ii}^k > 0 \quad \text{if Eq. (7) holds,} \quad (8b)$$

$$\sum_{j=1}^{N} \tilde{P}_{ij}^k = 1 \quad (8c)$$
Rewrite Eq. (6) as

$$g(n) = \max_{k \in \mathbb{N}} \left( \frac{q_N^k}{T_N^k} \right) \tau + \sum_{j=1}^{N-1} \hat{P}_{Nj}^k v_j(n),$$

(9a)

$$v_i(n+1) = \max_{k \in \mathbb{N}} \left( \frac{q_i^k}{T_i^k} \right) \tau - g(n) \tau + \sum_{j=1}^{N-1} \hat{P}_{ij}^k v_j(n)$$

(9b)

These are identical in form with Eq. (4) and therefore the scheme will converge if every $\hat{P}^A$ is both unichained and aperiodic. These two conditions are consequences of Eqs. (8a) and (8b), respectively. (Further, Eq. (8a) shows that the unique subchains of $P^A$ and $\hat{P}^A$ agree). The maximal-gain policies for Eqs. (1) and (9) can be shown to agree, so these conditions need hold only for maximal-gain policies [10].

6. Motivating the Iterative Scheme

Examination of the divergent example below Eqs. (5a and 5b) suggests that instability arises from faulty estimates $g(n)$ of $g^*$. Better estimates can be obtained if the process is sampled more frequently, say at small intervals $\tau$. A second advantage of equally-spaced samplings, of course, is that White's scheme may be employed.

Since only the mean holding times are required, there is no loss of generality in pretending that the original process, while in state $i$ with alternative $k$ active, earns rewards at a constant rate $q_i^k/T_i^k$ and has exponentially-distributed holding times with mean $T_i^k$. The constant transition rate to state $j$ (including the possibility $j = i$) is therefore $\hat{P}_{ij}^k/T_i^k$.

Since the system appears the same at any instant while in state $i$ as it does upon entering state $i$, decisioning at transitions can be replaced by decisioning at every instant. The latter can be replaced by decisioning at small intervals $\tau$. This leads to a Markov decision process with one step rewards $(q_i^k/T_i^k) \tau$ and transition probabilities $\hat{P}_{ij}^k = \left( \frac{P_{ij}^k}{T_i^k} \right) \tau$ for $j \neq i$. White's equations for this process are precisely Eq. (9a and 9b).

7. Bounds on $g^*$

Odoni [11] has shown how White's scheme provides upper and lower bounds on $g^*$. The equivalence between our Eqs. (9a and 9b) and Odoni's Eqs. (5 and 6) is through identification of our $g(n) \tau$ with $y_n(n)$ and $v_i(n)$.
with $W_i(n), 1 \leq i \leq N - 1$. Applied to Eqs. (6a and 6b), Odoni's results show that

$$g^+(n) = \max \left[ g(n), \max_{1 \leq i \leq N - 1} \max_{k \in x_i} \left[ \frac{q_i^k + \sum_{j=1}^{N-1} P_{ij}^k v(n)_j - v(n)_i}{T_i^k} \right] \right]$$

is monotone decreasing to $g^*$, while

$$g^-(n) = \min \left[ g(n), \min_{1 \leq i \leq N - 1} \max_{k \in x_i} \left[ \frac{q_i^k + \sum_{j=1}^{N-1} P_{ij}^k v(n)_j - v(n)_i}{T_i^k} \right] \right]$$

is monotone increasing to $g^*$. For large $n$, these provide tight bounds on $g^*$.

8. CHOICE OF $\tau$

Equations (6a and 6b) are a finite-difference approximation to the differential equations

$$\frac{dv_i}{dt} = \max_{k \in x_i} \left[ \frac{q_i^k + \sum_{j=1}^{N-1} P_{ij}^k v_j - v_i}{T_i^k} \right] - \max_{k \in x_i} \left[ \frac{q_i^k + \sum_{j=1}^{N-1} P_{ij}^k v_j}{T_i^k} \right]$$

1 \leq i \leq N - 1 \quad (10)

(which the above analysis shows to be asymptotically stable in the large with a unique equilibrium point) and the choice of proper step size $\tau$ is a classical problem: the finite-difference scheme converges too slowly if $\tau$ is too small, and not at all if $\tau$ is too large. In general, convergence of Eqs. (6a and 6b) is swiftest if $\tau$ is chosen neither infinitesimal nor as large as Eq. (7) permits. To show this, consider again the 2-state example with one alternative per state. Eqs. (6a and 6b) become

$$v(n + 1)_i = a + (1 - b) v(n)_i \quad (11)$$

where

$$a = \tau \left( \frac{q_1}{T_1} - \frac{q_2}{T_2} \right) \quad \text{and} \quad b = \tau \left( \frac{P_{12}}{T_1} + \frac{P_{21}}{T_2} \right).$$

Using Eq. (7), $b$ must lie between $0^+$ and (at most) $2^-$.

This 2-state example shows, incidentally, that Eqs. (6a and 6b) will not generally converge, not even when every $P^A$ is unichained and aperiodic, if the strict inequality in Eq. (7) is relaxed to $\leq$. This follows from the non-
convergence of Eq. (11) when \( b = 2 \), contrasted with a contrived 2-state example where the relaxed Eq. (7) permits \( b = 2 \). Consequently the bound in Eq. (7) is generally the best possible. This may be of independent interest because Eq. (10) provides a rare instance of an asymptotically stable set of nonlinear differential equations for which one can explicitly state the range of step sizes for which the finite-difference approximation is also asymptotically stable.

REFERENCES