Forward recursion for Markov decision processes with skip-free-to-the-right transitions, part I : theory and algorithm

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 11, No. 2, May 1986
Printed in U.S.A

FORWARD RECURSION FOR MARKOV DECISION PROCESSES WITH SKIP-FREE-TO-THE-RIGHT TRANSITIONS, PART I: THEORY AND ALGORITHM*

JACOB WIJNGAARD†‡ AND SHALER STIDHAM, JR.§**

We consider a Markovian decision process with countable state space (states $0, 1, 2, \ldots$) which is skip-free to the right (a transition from i to j is impossible if $j > i + 1$). In this type of system it is easy to calculate by forward recursion the maximal total expected reward going from state 0 to state i; the same can be done, of course, for the case where a constant g is subtracted from the one-period reward function (g-revised reward). Let $-w^g(i)$ be the maximal total expected g-revised reward going from state 0 to state i. We show that $w^g(\cdot)$ satisfies the average-reward optimality equation. If $w^g(\cdot)$ satisfies a growth condition, then $g = g^*$, the maximal average reward. For all other g, the function w^g increases or decreases so fast that this cannot be the case. Thus, in principle the solution w^g can be used to check if $g < g^*$ or $g > g^*$, which suggests a method for approximating g^* and an associated average-return optimal policy. We develop an efficient algorithm based on this idea. In a companion paper we shall show how the algorithm, or modifications of it, can be applied to some special cases, such as control of arrivals to a queue, control of the service rate, and controlled random walks.

We consider a semi-Markov decision process (Ross [7]) on a countable state space $S = \{0, 1, 2, \ldots\}$. Let $A(i)$ be the set of possible actions in state i and for $a \in A(i)$ let the expected time and the expected reward until the next transition be denoted by $\tau(i,a)$ and $r(i,a)$, respectively. The transition probabilities from state i under action $a \in A(i)$ are denoted $p_{ij}(a)$, $j \in S$. We assume that the following conditions are satisfied.

Condition 1. The system is skip-free to the right (Keilson [2], [3]) and nondegenerate: $p_{ij}(a) = 0$, for all $j > i + 1$ and all $a \in A(i)$, and $p_{i,i+1}(a) > 0$, for all i and all $a \in A(i)$.

Condition 2. For each state i the set of possible actions $A(i)$ can be represented by a compact metric space such that $\tau(i,a)$, $r(i,a)$, and $p_{ij}(a)$ are continuous in $a \in A(i)$.

Our objective is to construct an efficient algorithm, exploiting the right-skip-free transition structure, for finding the maximal long-run average reward, g^*. The average-reward optimality equation for this problem is:

$$x(i) = \max_{a \in A(i)} \left\{ r(i,a) - g\tau(i,a) + \sum_{j=0}^{i+1} p_{ij}(a) x(j) \right\}, \qquad i \geqslant 0. \tag{1}$$

(Condition 2 ensures that the max is attained for each state i.) An equivalent form of

*Received June 13, 1983; revised March 4, 1985.
AMS 1980 subject classification. Primary: 9C347. Secondary: 90C39.
IAOR 1973 subject classification. Main: Programming:dynamic. Cross references: Markov Processes.
OR/MS Index 1978 subject classification. Primary: 117 Dynamic programming/Markov.
Key words. Markov decision process; dynamic programming; skip-free transitions; forward recursion.

† Eindhoven University of Technology.

‡ The research of this author was begun during a visit to the Program in Operations Research, North Carolina State University.

§ North Carolina State University.

** The research of this author was partially supported by the National Science Foundation, Grant ENG 78-24420, and the U.S. Army Research Office, Contract DAAG29-82-K-0152, at North Carolina State University, Raleigh, NC, USA, and by the Science and Engineering Research Council, at the University of Cambridge, UK.

0364-765X/86/1102/0295$01.25

this equation is:

$$x(i+1)-x(i)=\min_{a\in A(i)}\left[p_{i,i+1}(a)\right]^{-1}\left\{g\tau(i,a)-r(i,a)+\sum_{j=0}^{i-1}p_{ij}(a)\left[x(i)-x(j)\right]\right\}. \tag{2}$$

(A formal proof of this equivalence can be constructed along the same lines as the proof of Lemma 5 in §2 of this paper.) It is clear that for each g setting $x(0)=0$ fixes a solution to (2): call this solution $w^g(\cdot)$. In fact (see Lemma 5) $-w^g(i)$ equals the maximal total expected reward if the process is started in state 0 and stopped at the first entrance into state i (before earning a reward) and the reward function is $r(\cdot,\cdot)-g$ instead of $r(\cdot,\cdot)$ (*g-revised reward*).

Let X_n denote the state at stage n. If the expected value of $w^g(X_n)$, given $X_0=i$, is $o(n)$ as $n\to\infty$, for each control strategy and each starting state i, then $g=g^*$ (Ross [7, p. 145]). For all other g the function w^g increases or decreases so fast that this cannot be the case. Thus, in principle, the solution w^g can be used to check if $g<g^*$ or $g>g^*$, which suggests a method for approximating g^* and an associated average-reward optimal strategy. Rather than work with the above limiting condition on w^g, we formulate an alternative condition that is more easily checked and equivalent to the above condition when the reward and transition probabilities satisfy certain regularity conditions.

The idea of solving an average-reward Markovian decision process by approximating g^* and solving a sequence of stopping problems with g-revised rewards has been exploited by other researchers (see, e.g., Low [4], [5], Miller [6]). The difference between Miller's approach and ours is that, roughly speaking, Miller checks whether $g=g^*$ by looking at $x(0)$, whereas we look at $x(i)$ for large i. In addition, our method makes essential use of the right-skip-free transition structure.

§1 of this paper lays the groundwork by providing an appropriate formulation of the maximal average-reward problem. We give conditions which, together with Conditions 1 and 2 above, are sufficient for an average-optimal strategy to exist and for $v^g(i)$, the optimal total expected g-revised reward going from state i to state 0, to exist and to satisfy the average-reward optimality equation with $g=g^*$. Our conditions are not stringent and seem to be satisfied, for example, in most queueing-control applications. The proofs of the results in this section are rather technical and are therefore deferred to an appendix.

§2 develops the algorithm for approximating g^*, based on the average-reward model of §1. The algorithm estimates g^* by a value g, calculates the associated values of $w^g(i)$, $i\geqslant 0$, by forward recursion from (2), and then checks to see if w^g coincides with v^g. Since this will be true if and only if $v^g(0)=0$ and the latter is true if and only if $g=g^*$, we have thereby a mechanism for checking whether or not $g=g^*$. The check on whether w^g coincides with v^g requires additional conditions on the rewards and transition probabilities: specifically, that the rewards are "almost polynomial" (Condition 7) and that the chain has a "uniform tendency to the left" from large enough states (Condition 8). Under these conditions, it follows (from results proved in an appendix) that v^g is the unique solution, in the class of almost polynomial functions, to the functional equation for the total expected g-revised reward until state 0 is reached. Moreover, $w^g=v^g$ if and only if w^g is also in this class. We are able to develop an algorithm, based on upper and lower bounds on v^g, for checking whether or not w^g belongs to this class.

Finally, in §3 we show how problems with discounted rewards can also be solved by our algorithm, by first replacing the discounted-reward problem by an equivalent average-reward problem.

In a companion paper we shall show how the algorithm can be applied in some special cases, including control of arrivals to a queue, control of the service rate, and controlled random walks. Modifications of the algorithm to cope with nonstandard applications will be presented. We shall also discuss alternative procedures for checking whether or not $w^g = v^g$, based on relations involving the stationary distribution of the chain. In the special case of a controlled random walk, the conditions based on the stationary distributions are shown to coincide with the original conditions.

1. Average-reward semi-Markov decision process. Our framework for average-reward maximization in the semi-Markov decision process introduced in the previous section will restrict attention to the set $\mathscr{A}$ of stationary strategies α. For each state $i \in S$, $\alpha(i) \in A(i)$ is the action dictated by strategy α whenever the process is in state i. For each α, let P_α denote the transition-probability matrix and let r_α and τ_α denote the reward and transition-time functions, respectively. That is, P_α is a matrix with i-jth component equal to $p_{ij}(\alpha(i))$ and r_α (τ_α) is a vector with ith component equal to $r(i, \alpha(i))$ $(\tau(i, \alpha(i)))$.

We shall introduce some conditions, which, together with Conditions 1 and 2, ensure that a strategy exists that maximizes long-run average reward and that it can be found from the average-reward optimality equation. Proofs of all the results in this section are given in Appendix 2.

Condition 3. There exist positive numbers L, M_1, and M_2 such that $L \leqslant \tau(i,a) \leqslant M_1$, and $r(i,a) \leqslant M_2$, for all i, $a \in A(i)$.

Condition 4. For each i there is a positive number ϵ_i and an $n > i$ such that, starting in i, the probability that the first visit to 0 occurs before the first visit to $\{n+1, n+2, \dots\}$ is at least ϵ_i, independent of the strategy.

Note that Condition 4 is more or less naturally satisfied in most queueing-control applications. It follows from Condition 4 that there exists a B_n such that the expected number of visits to $[1, n-1]$ before the next visit to state 0, starting in state i and following strategy α, is bounded above by B_n, for all $i \geqslant 0$ and all α.

Before stating the final two conditions, we introduce some additional useful notation. For each strategy α and each integer $n \geqslant 0$, we define an operator $P_{\alpha n}$ on the set of functions $f\colon S \to R$ by

$$(P_{\alpha n} f)(i) := \sum_{j=n}^{i+1} p_{ij}(\alpha(i)) f(j), \qquad i \geqslant 0.$$

Note that $P_{\alpha n} f$ can be interpreted as the (column) vector resulting from pre-multiplying the (column) vector f by the matrix P_α, after replacing columns $0, 1, \dots, n-1$ of P_α by columns of zeroes. Also, when it is well defined, the quantity $\sum_{t=0}^\infty (P_{\alpha n}^t f)(i)$ is the expected total accumulation of f until the next visit to $[0, n-1]$, starting in state i and following strategy α. We shall be particularly interested in the case where $n = 1$ and $f = r_\alpha$ or $f = \tau_\alpha$. The quantities $\sum_{t=0}^\infty (P_{\alpha 1}^t r_\alpha)(i)$ and $\sum_{t=0}^\infty (P_{\alpha 1}^t \tau_\alpha)(i)$ are, respectively, the expected total reward and expected total time until the next visit to state 0, starting in state i and following strategy α. In Condition 5 we require the finiteness of both these quantities for at least one α.

Condition 5. There is a strategy $\hat{\alpha}$ such that $\sum_{t=0}^\infty (P_{\hat{\alpha} 1}^t r_{\hat{\alpha}})(i)$ and $\sum_{t=0}^\infty (P_{\hat{\alpha} 1}^t \tau_{\hat{\alpha}})(i)$ are both finite for all i.

For each α we define the average reward, starting in i, as follows (cf. Ross [7, p. 159]:

$$g_\alpha(i) := \limsup_{n \to \infty} \left[\sum_{t=0}^{n-1} (P_\alpha^t r_\alpha)(i) \right] \Big/ \left[\sum_{t=0}^{n-1} (P_\alpha^t \tau_\alpha)(i) \right].$$

Condition 5 implies (see, e.g., Ross [7, Theorem 7.5]) that, for $\alpha = \hat{\alpha}$, $g_\alpha(i)$ is independent of i and takes the form

$$g_\alpha(i) = g_\alpha := \left[\sum_{t=0}^{\infty} (P_{\alpha 1}^t r_\alpha)(0) \right] \Big/ \left[\sum_{t=0}^{\infty} (P_{\alpha 1}^t \tau_\alpha)(0) \right]. \tag{3}$$

Our next and last condition will make it possible to restrict attention to strategies α for which (3) holds, without loss of optimality.

Condition 6. There exist a real number g_0 and an integer n_0, such that $g_0 < g_{\hat{\alpha}}$ and $r(i,a) \leqslant g_0 \tau(i,a)$, for all $i > n_0$, $a \in A(i)$.

For the remainder of this section we assume that Conditions 1–6 hold.

THEOREM 1. *Let α be a strategy for which $g_\alpha(i) \geqslant g_{\hat{\alpha}}$ for all i. Then the sums $\sum_{t=0}^{\infty}(P_{\alpha 1}^t r_\alpha)(i)$ and $\sum_{t=0}^{\infty}(P_{\alpha 1}^t \tau_\alpha)(i)$ are both finite for all i.*

It follows from Theorem 1 that (3) holds for any strategy that has $g_\alpha(i) \geqslant g_{\hat{\alpha}}$ for all i. The problem of finding an optimal strategy has thus been reduced to finding a strategy that maximizes the expression g_α given in (3), among strategies for which both numerator and denominator are finite.

The next step is to construct a solution to the average-return optimality equation, from which an optimal policy can be determined. To this end, we first consider the problem of maximizing the total expected g-revised reward until the next visit to state 0 from each starting state i. For arbitrary g, i, and α, define $v_\alpha^g(i)$ as the total expected accumulation of $r_\alpha - g\tau_\alpha$ until the next visit to 0, starting in i and following strategy α. Let $v^g(i) := \sup_{\alpha \in \mathcal{A}} v_\alpha^g(i)$. Under Conditions 1–6, an argument like that used in the proof of Theorem 1 shows that $v^g(i)$ is well defined and satisfies the following functional equation in $x(\ )$:

$$x(i) = \max_{a \in A(i)} \left\{ r(i,a) - g\tau(i,a) + \sum_{j=1}^{i+1} p_{ij}(a) x(j) \right\}. \tag{4}$$

Condition 5 implies that $v^g(i)$ is finite for all i. Moreover, a strategy that takes a maximizing action in each state i is optimal for this problem (cf. Schäl [8]).

LEMMA 2. *Let α be a strategy for which both the expected total reward and the expected total time until the next visit to state* 0, *starting in i, are finite. Then $\lim_{n\to\infty}(P_{\alpha 1}^n v^g)(i) = 0$ for all g, i.*

LEMMA 3. *$v^g(i)$ is continuous in $g \geqslant g_0$ for each state i.*

By Condition 3 it is clear that for $g = M_2/L$, we have $v^g(0) \leqslant 0$. On the other hand, for $g = g_{\hat{\alpha}}$, we have $v^g(0) \geqslant 0$ (see Condition 5). The existence of a g^* such that $v^{g^*}(0) = 0$ now follows from Lemma 3. Hence for this g^* we have ($i \geqslant 0$)

$$v^{g^*}(i) = \max_{a \in A(i)} \left\{ r(i,a) - g^* \tau(i,a) + \sum_{j=0}^{i+1} p_{ij}(a) v^{g^*}(j) \right\}, \tag{5}$$

that is, $(g^*, v^{g^*}(\cdot))$ satisfy the average-reward optimality equation (1).

The next theorem asserts that g^* is the maximal average reward and that g^* is realized by α_{g^*}, where α_{g^*} is a strategy that maximizes the right-hand side of (5).

THEOREM 4. *The average reward under α_{g^*} equals g^*. There are no strategies in $\mathcal{A}$ with average reward greater than g^*.*

2. Algorithm.

Recall the average-reward optimality equation (1). From the previous section we know that $v^g(\cdot)$ satisfies (1) if and only if $v^g(0) = 0$, in which case $g = g^*$. But we remarked in the introductory section that, for any g, fixing $x(0) = 0$

generates a unique solution to (1), denoted $w^g(\cdot)$, which can be calculated recursively from (2). Therefore, $g = g^*$ if and only if $v^g(\cdot) = w^g(\cdot)$. We shall use this observation as the basis of an algorithm for iterative approximation of g^*. The following lemma gives an interpretation of $w^g(\cdot)$ which is useful in understanding the algorithm.

LEMMA 5. *Assume Conditions* 1 *and* 2. *For* $0 \leqslant i \leqslant j$ *let* $z^g(i,j)$ *be the supremum of the total expected accumulation of* $r - g\tau$ *until the first visit to state* j, *starting in state* i ($z^g(i,i) = 0$). *Then* $z^g(0,i) = -w^g(i)$, *so that* $w^g(i)$ *is the infimum of the total expected accumulation of* $g\tau - r$ *until the first visit to state* i, *starting in state* 0. *The infimum is attained for some stationary strategy.*

PROOF. By Conditions 1 and 2 the probability $p_{i,i+1}(a)$ is bounded away from 0. Thus $z^g(\cdot,\cdot)$ is well defined and finite. We may write

$$z^g(i,i+1) = \max_{a \in A(i)} \left\{ r(i,a) - g\tau(i,a) + \sum_{j=0}^{i} p_{ij}(a) z^g(j,i+1) \right\}. \tag{6}$$

The skip-free characteristic implies $z^g(j,i+1) = z^g(j,i) + z^g(i,i+1)$ for $j \leqslant i$. Substituting this into (6) yields

$$p_{i,i+1}(a) z^g(i,i+1) \geqslant r(i,a) - g(i,a) + \sum_{j=0}^{i} p_{ij}(a) z^g(j,i)$$

for all $a \in A(i)$, with equality for at least one a. Thus, using $z^g(j,i) = z^g(0,i) - z^g(0,j)$ for $j \leqslant i$, we obtain

$$z^g(0,i+1) - z^g(0,i) = \max_{a \in A(i)} \left[p_{i,i+1}(a)\right]^{-1} \Bigg\{ r(i,a) - g\tau(i,a)$$

$$+ \sum_{j=0}^{i} p_{ij}(a)\left[z^g(0,i) - z^g(0,j)\right] \Bigg\}.$$

From equation (2) and the definition of $w^g(\cdot)$ it then follows that $z^g(0,i) = -w^g(i)$. It is also clear that a strategy that chooses a maximizing action in each state i is optimal. ∎

Throughout the remainder of this section, we shall assume that Conditions 1–6 hold, as well as the following two conditions:

Condition 7. The reward $r(i,a)$ is *almost polynomial*: $|\max_{a \in A(i)} r(i,a)|/\rho^i$ is bounded, for all $\rho > 1$.

Condition 8. There exist an $\epsilon > 0$ and positive integers N and k, with $k < N$, such that for all $i \geqslant N - 1$,

$$-p_{i,i+1}(a) + \sum_{j=i-k+1}^{i} (i-j) p_{ij}(a) + \sum_{j=0}^{i-k} k p_{ij}(a) \geqslant \epsilon.$$

In particular, Condition 8 implies that, under every strategy α, the Markov chain P_α has a *uniform tendency to the left*: $\sum_{j=0}^{i+1}(i-j)p_{ij}(\alpha(i)) \geqslant \epsilon$, $i \geqslant N-1$. That is, at every transition from a state $i \geqslant N - 1$, the expected jump to the left is uniformly positive.

These conditions will give us a way of checking whether or not $v^g = w^g$, and hence whether or not $g = g^*$. We shall show that Conditions 1–8 imply that $v^g(i)$ is almost polynomial and that $w^g(i)$ is almost polynomial if and only if $g = g^*$. Our algorithm exploits these facts by first computing upper and lower bounds on $v^g(i)$ that are almost polynomial, and then checking to see if $w^g(i)$ lies within these bounds. The next theorem establishes the basis for this approach and also indicates how to distinguish between the cases $g < g^*$ and $g > g^*$.

THEOREM 6. *For all g, the function v^g is almost polynomial; that is, $v^g(i)/\rho^i$ is bounded for all $\rho > 1$. For $g < g^*$ there is a $\rho > 1$ such that $w^g(i)/\rho^i$ is not bounded from below ($w^g(i)$ goes exponentially fast to $-\infty$). For $g > g^*$ there is a $\rho > 1$ such that $w^g(i)/\rho^i$ is not bounded from above ($w^g(i)$ goes exponentially fast to $+\infty$).*

PROOF. First note that Conditions 1–8 imply that Conditions A1.1, A1.2, and A1.3 of Appendix 1 hold for each fixed strategy α. The first assertion of the theorem then follows directly from Theorem A1.3 of Appendix 1, applied to the Markov chain generated by the strategy that yields v^g. To prove the second and third assertions, we define $w_\alpha^g(i)$ as the total expected accumulation of $g\tau_\alpha - r_\alpha$ until the first visit to i, starting in 0 and following strategy α. By Lemma 5, we have $w_\alpha^g(i) \geqslant w^g(i)$ for all i and all strategies. Let α_g be the strategy found in the construction of w^g. For $g < g^*$ we have

$$w^g = w_{\alpha_g}^g \leqslant w_{\alpha_{g^*}}^g < w_{\alpha_{g^*}}^{g^*} = w^{g^*} = v^{g^*}.$$

The difference $w_{\alpha_{g^*}}^{g^*}(i) - w_{\alpha_{g^*}}^g(i)$ is equal to the total expected accumulation of $(g^* - g)\tau$ until the first visit to i, starting in 0 and following strategy α_{g^*}. From Appendix 1 (see remarks in the final paragraph) it follows that this difference goes exponentially fast to $+\infty$. Hence, $w^g(i)$ goes exponentially fast to $-\infty$. For $g > g^*$ we have

$$w^g = w_{\alpha_g}^g > w_{\alpha_g}^{g^*} \geqslant w_{\alpha_{g^*}}^{g^*} = w^{g^*} = v^{g^*}.$$

As in the previous case we can now prove that $w^g(i)$ goes exponentially fast to $+\infty$. ∎

Next we derive upper and lower bounds for $v^{g^*}(i)$ which are almost polynomial.

First a lower bound. Take some strategy α. Then $v^{g^*}(i) \geqslant v_\alpha^{g^*}(i)$, which is the total expected accumulation of $r - g^*\tau$ until the next visit to 0, starting in state i and following strategy α. From Condition 3 we have $g^* \leqslant M_2/L$ and

$$r_\alpha(i) - g^*\tau_\alpha(i) \geqslant r_\alpha(i) - [M_2/L]\tau_\alpha(i).$$

The total expected accumulation of $r_\alpha(i) - [M_2/L]\tau_\alpha(i)$ until the next visit to 0, starting in i and following strategy α, can be calculated by the methods described in Appendix 1. This gives a lower bound for v^{g^*}.

Next an upper bound. First we need a lower bound for g^*. One possibility is to use the constant g_0 from Condition 6. A better lower bound might be $g_{\hat{\alpha}}$, where the strategy $\hat{\alpha}$ is from Condition 5, provided $g_{\hat{\alpha}}$ is easy to calculate. Alternatively, one can make a guess g and then calculate w^g and check whether it violates the lower bound for v^{g^*} we have already. Suppose g is a lower bound with $g_0 \leqslant g \leqslant g^*$. Then

$$r(i,a) - g^*(i,a) \leqslant r(i,a) - g\tau(i,a).$$

Choose n_0 such that $r(i,a) - g(i,a) < 0$ for $i > n_0$ and for all $a \in A(i)$ (cf. Condition 6). Define $r'(i,a)$ in the following way (with $K := L$ if $g > 0$, and $K := M_1$ if $g < 0$):

$$r'(i,a) := M_2 - gK, \qquad \text{for} \quad i \leqslant n_0,$$
$$r'(i,a) := 0, \qquad \text{for} \quad i > n_0.$$

Then $r(i,a) - g^*\tau(i,a) \leqslant r'(i,a)$. For some $n > n_0$, let B_n be an upper bound for the expected number of visits to $[1,n]$ until the next visit to 0, starting in i (see remarks following Condition 4 in §1). Then an upper bound for $v^{g^*}(i)$ is $(M_2 - gK)B_n$. Using the ϵ_i in Condition 4, it is possible to derive an upper bound B_n in an analytical way.

It is now possible to sketch the algorithm to approximate g^*.

Step 1. Calculate a lower bound and an upper bound for v^{g^*}.

Step 2. Choose g_1 equal to the lower bound for g^* used in the calculation of the upper bound for v^{g^*}. Choose g_2 equal to M_2/L. Then $g_1 \leqslant g^* \leqslant g_2$.

Step 3. Calculate w^g for $g := (g_1 + g_2)/2$. If w^g violates the lower bound for v^{g^*} then $g < g^*$ and we define $g_1 := g$. If w^g violates the upper bound for v^{g^*} then $g > g^*$ and we define $g_2 := g$.

Step 4. If $g_2 - g_1$ is greater than some fixed $\epsilon > 0$, then go back to Step 3. Otherwise the iteration stops and $(g_1 + g_2)/2$ is chosen as the approximation for g^*.

The algorithm gives a good approximation for g^*. The following corollary to Theorem 6 shows that it also yields a good strategy.

COROLLARY 7. *Let $g < g^*$. Let α_g be the strategy found in constructing w^g. The average reward under α_g is at least equal to g.*

PROOF. Let $\tilde{g}$ be the average reward under α_g. Then $v_{\alpha_g}^{\tilde{g}} = w_{\alpha_g}^{\tilde{g}}$ and $w_{\alpha_g}^{\tilde{g}}$ is almost polynomial. Since $w_{\alpha_g}^{g}$ goes exponentially fast to $-\infty$, it follows that $g < \tilde{g}$ (cf. proof of Theorem 6). ∎

Applying Corollary 7 to $g = g_1 < g^*$ yields the desired result.

The algorithm can be used for all semi-Markov decision processes satisfying Conditions 1–8, but its performance depends rather heavily on certain characteristics of the decision process. Most important is the strength of the tendency to the left. The rate of divergence of w^g for $g < g^*$ is determined by the strength of the tendency to the left of the strategy α_{g^*}, which is the average-reward optimal strategy (see proof of Theorem 6). For most problems this tendency to the left will be rather strong and therefore the check $g < g^*$ will work rather well in general. The rate of divergence of w^g for $g > g^*$ is determined by the strength of the tendency to the left of α_g, which is the strategy found in constructing w. We may expect the tendency to the left of α_g to be weaker than that of α_{g^*}. This can be seen as follows. The quantity $w^g(i)$ is the minimal expected total accumulation of $g\tau - r$ until the first visit to i, starting in 0. If $g > g^*$ we know that on the average r cannot keep up with $g\tau$. Therefore, to keep the total expected accumulation of $g\tau - r$ until the first visit to i small, it is necessary to keep the time until the first visit to i small. This implies a kind of "minimal" tendency to the left for α_g. As long as all strategies have a strong left-tendency, there is no problem. Otherwise, there may be numerical difficulties.

In such cases it is possible to use α_g rather than w^g to check whether $g > g^*$. For example, one may add a decision $\tilde{a}$ such that $p_{i,i+1}(\tilde{a}) = 1$. One must choose the reward and transition time of this action so that it is certainly not used under an optimal strategy. In general this is not difficult since an optimal strategy has a tendency to the left in most cases. Now the check on $g < g^*$ is executed as before by calculating w^g. The existence of the extra decision $\tilde{a}$ does not interfere with this step, since only the tendency to the left of α_{g^*} is used here. The check on $g > g^*$, however, is executed by considering α_g. If we find that, for some i, $\alpha_g(i) = \tilde{a}$, then we know that $g > g^*$. For a special queueing-control problem described in Part II, this idea of using α_g for checks on $g > g^*$ is exploited. There it is worked out more precisely.

Another condition which may be too strong is the requirement that $p_{i,i+1}(a) > 0$ for all i, $a \in A(i)$ (see Condition 1). In problems involving control of arrivals to a queue, it is indeed unnatural to have this condition, since an optimal policy may reject all arrivals in some states. We consider arrival-control problems in Part II, where we show how problems caused by $p_{i,i+1}(a) = 0$ can be circumvented.

3. Discounted rewards.

We have developed the theory and the algorithm for the average-reward criterion. Problems with discounted reward can also be solved by transforming them into equivalent average-reward problems (cf. Derman [1, p. 115]).

We illustrate this transformation for the Markov case, but it can also be done for the semi-Markov case.

Let $p_{ij}(a)$, $r(i,a)$ be the transition probabilities and reward function for a discounted-reward Markov decision process with state space $\{0,1,2,\dots\}$, infinite horizon, and one-stage discount rate β. Consider an average-reward Markov decision process with extended state space $\{-1,0,1,2,\dots\}$ and transition probabilities and reward function as follows:

$$\begin{aligned} p'_{ij}(a) &:= \beta p_{ij}(a), && \text{for } i,j=0,1,2,\dots, \quad a\in A(i);\\ p'_{i,-1}(a) &:= 1-\beta, && \text{for } i=0,1,2,\dots, \quad a\in A(i);\\ r'(i,a) &:= r(i,a), && \text{for } i=0,1,2,\dots, \quad a\in A(i). \end{aligned}$$

In state -1 there is only one possible action, a', and

$$p'_{-1,-1}(a') := 1-\beta; \qquad p'_{-1,0}(a') := \beta; \qquad r'(-1,a') := 0.$$

If the original (discounted-reward) decision process is skip free to the right, then so is the new (average-reward) decision process.

Using the methods developed in §§1 and 2, one may find an average-reward optimal strategy for the new problem and an associated solution $v'(\cdot)$ of the average-reward optimality equation, with $v'(-1)=0$. Thus, for $i=-1,0,1,2,\dots$

$$v'(i) = \max_{a\in A(i)}\left\{ r'(i,a) - g^* + \sum_{j=0}^{i+1} p'_{ij}(a)v'(j)\right\}.$$

Substitution for $r'(i,a)$ and $p'_{ij}(a)$ yields, for $i=0,1,2,\dots,$

$$v'(i) = \max_{a\in A(i)}\left\{ r(i,a) - g^* + \sum_{j=0}^{i+1} \beta p_{ij}(a)v'(j)\right\}. \tag{7}$$

Define $v(i) := v'(i) + g^*/(1-\beta)$, for $i=0,1,2,\dots$. Then

$$v(i) = \max_{a\in A(i)}\left\{ r(i,a) + \beta \sum_{j=0}^{i+1} p_{ij}(a)v(j)\right\},$$

and the maximizing a is the same as in (7). Hence the restriction of the optimal strategy for the average-reward problem to the states $i=0,1,2,\dots,$ is an optimal strategy for the discounted-reward problem and the solutions to the two optimality equations differ only by a constant.

Appendix 1. Total expected reward in a right-skip-free Markov chain. In this appendix we derive some results that yield information about the behavior of the total expected g-revised reward until the next visit to state 0, following a particular strategy α, as a function of the starting state i. We establish conditions under which this function is almost polynomial and the unique solution to a functional equation analogous to (4), in the class of almost polynomial functions. This result is applied (in Theorem 6 in §2) to form the basis for our algorithm for approximating the maximal average reward g^*.

We consider a Markov chain $P=(p_{ij})$ on the state space $S=\{0,1,2,\dots\}$, which is skip-free to the right and nondegenerate (cf. Condition 1). Let $r(i)$ denote the one-stage reward received when the chain is in state i, $i\geqslant 0$. (In our applications in §2, we might have, for example, $P=P_\alpha$, $r=r_\alpha - g\tau_\alpha$, with α a strategy that attains $v^{\mathscr{G}}$.)

For $\rho>1$ let $\mathscr{B}_\rho$ be the set of all complex-valued functions f on the state space S such that $|f(i)|/\rho^i$ is bounded in i. One can make such a set a Banach space by defining a norm $\|f\|_\rho := \sup_{i \geqslant 0}|f(i)|/\rho^i$. We shall use the following conditions (cf. Conditions 7 and 8 in §2).

Condition A1.1. For all $\rho > 1$, $r \in \mathscr{B}_\rho$. That is, r is almost polynomial.

Condition A1.2. There exist an $\epsilon > 0$ and positive integers N and k, with $k < N$, such that

$$-p_{i,i+1} + \sum_{j=i-k+1}^{i} (i-j)p_{ij} + \sum_{j=0}^{i-k} kp_{ij} \geqslant \epsilon, \qquad \text{for all} \quad i \geqslant N-1.$$

Let the operator P_n, $n \geqslant 0$, be defined in an analogous way to $P_{\alpha n}$ in §1.

LEMMA A1.1. *Assume Condition* A1.2. *Then there exists a* $\rho^* > 1$ *such that* P_N *is a contraction operator on* $\mathscr{B}_{\rho^*}$.

PROOF. Let $\rho > 1$ and $f \in \mathscr{B}_\rho$. Consider $P_N f$. For each $i \geqslant 0$, we have

$$|(P_N f)(i)|/\rho^i \leqslant \rho^{-i} \sum_{j=N}^{i+1} p_{ij}|f(j)| \leqslant \|f\|_\rho \sum_{j=N}^{i+1} p_{ij}\rho^{j-i}.$$

For $i < N-1$, the right-hand side of the last inequality equals 0. Therefore, to prove the lemma it suffices to show that there exists a δ, $0 \leqslant \delta < 1$, such that $\sum_{j=0}^{i+1} p_{ij}\rho^{j-i} \leqslant \delta$, for $\rho > 1$ and sufficiently close to 1, for all $i \geqslant N-1$. Now

$$\begin{aligned}
\sum_{j=0}^{i+1} p_{ij}\rho^{j-i} &\leqslant p_{i,i+1} + \sum_{j=i-k+1}^{i} p_{ij}\rho^{j-i} + \sum_{j=0}^{i-k} p_{ij}\rho^{-k} \\
&\leqslant 1 - (\rho-1)\left[-p_{i,i+1} + \sum_{j=i-k+1}^{i} p_{ij}(i-j)\rho^{j-i-1} + \sum_{j=0}^{i-k} p_{ij}k\rho^{-k-1}\right] \\
&\leqslant 1 - (\rho-1)\left[\rho^{-k-1}\left\{-p_{i,i+1} + \sum_{j=i-k+1}^{i} (i-j)p_{ij} + \sum_{j=0}^{i-k} kp_{ij}\right\}\right. \\
&\qquad\qquad \left. - (1-\rho^{-k-1})p_{i,i+1}\right] \\
&\leqslant 1 - (\rho-1)\left[\rho^{-k-1}(1+\epsilon) - 1\right].
\end{aligned}$$

The second inequality follows from the fact that the expression on the right-hand side of the first inequality is convex in $\rho \geqslant 1$. The fourth inequality follows from Condition A1.2. Now choose $\rho = \rho^* > 1$ such that the term in brackets is > 0. In this case the right-hand side of the last inequality is < 1, so that there exists a $\delta < 1$ such that $\sum_{j=0}^{i+1} p_{ij}(\rho^*)^{j-i} \leqslant \delta < 1$, for all $i \geqslant N-1$, and hence P_N is a contraction on $\mathscr{B}_{\rho^*}$. ∎

Let $f \in \mathscr{B}_{\rho^*}$. It follows from Lemma A1.1 that $\sum_{t=0}^{\infty} P_N^t f$ is also an element of $\mathscr{B}_{\rho^*}$, if Condition A1.2 holds. The value $\sum_{t=0}^{\infty}(P_N^t f)(i)$ may be interpreted as the total expected accumulation of f until the next visit to $[0, N-1]$ from starting state i. (Note that for $i < N-1$, $\sum_{t=0}^{\infty}(P_N^t f)(i) = f(i)$.)

Our goal is to study the behavior of the total expected reward until the next visit to state 0, as a function of the starting state i. To this end we consider the imbedded Markov chain on $[0, N-1]$. Let $R(i)$ denote the one-stage reward function for this chain, so that $R(i) = \sum_{t=0}^{\infty}(P_N^t r)(i)$, which belongs to $\mathscr{B}_{\rho^*}$ if Conditions A1.1 and A1.2 hold. The transition operator associated with the imbedded chain is denoted by Q and is defined on the set of functions f with support in $[0, N-1]$ by

$$(Qf)(i) := \sum_{t=0}^{\infty} (P_N^t P f)(i).$$

For any state i, $(Qf)(i)$ can be interpreted as the expected terminal reward received if the system is started in state i and stopped at the next visit to $[0, N-1]$, with a terminal reward $f(j)$ if termination occurs in state $j \in [0, N-1]$. For $i, j \in [0, N-1]$ and f the indicator of state j, $(Qf)(i)$ gives the probability of a one-step transition from i to j in the embedded chain on $[0, N-1]$. Let Q_1 be the restriction of Q to $[1, N-1]$, so that $Q_1 = \sum_{t=0}^{\infty} P_N^t P_1$. The following condition guarantees recurrence of state 0.

Condition A1.3. The operator Q_1 is an n-stage contraction with respect to the metric $\| \cdot \|_{\rho^*}$ on the space of functions with support in $[1, N-1]$.

This condition is satisfied, for example, if there is a positive probability of visiting state 0 before leaving $[1, N-1]$.

If Conditions A1.1, A1.2, and A1.3 are satisfied, one can write the total expected reward until the next visit to state 0, starting in state i, as

$$v(i) := \sum_{t=0}^{\infty} (Q_1^t R)(i), \qquad i \geqslant 0.$$

In this case it is easy to verify that v is an element of β_{ρ^*} and that v satisfies

$$x(i) = R(i) + (Q_1 x)(i). \tag{8}$$

Moreover, since $R(\cdot) \in \mathscr{B}_{\rho^*}$ and Q_1 is an n-stage contraction, v is the unique solution in $\mathscr{B}_{\rho^*}$ to (8). But, by standard arguments applied to the original Markov chain P, v must also satisfy the more familiar functional equation

$$x(i) = r(i) + \sum_{j=1}^{i+1} p_{ij} x(j), \qquad i \geqslant 0. \tag{9}$$

It is easy to show that, since P_N is a contraction on $\mathscr{B}_{\rho^*}$, any solution to (9) in $\mathscr{B}_{\rho^*}$ is also a solution to (8). We have thus proved

THEOREM A1.2. *Assume Conditions* A1.1, A1.2, *and* A1.3. *Then v is the unique solution in $\mathscr{B}_{\rho^*}$ to the functional equation* (9).

Equation (9) can be rewritten ($i \geqslant 0$) as

$$x(i+1) - x(i) = (p_{i,i+1})^{-1} \left[\sum_{j=0}^{i-1} p_{ij}(x(i) - x(j)) + p_{i0} x(0) - r(i) \right]. \tag{10}$$

This shows that each $x(0)$ generates a unique solution to the equation. An easy induction on i shows that the general solution to (10) takes the form $w + fx(0)$, where w is the solution generated by $w(0) = 0$, and f is the solution to the corresponding homogeneous equations with $f(0) = 1$. By an argument parallel to that used in the proof of Lemma 5 in §2, it can be shown that $-w(i)$ equals the total expected reward until the first visit to state i, starting in state 0.

The unique solution in $\mathscr{B}_{\rho^*}$ to the homogeneous equations is identically zero. Hence f cannot be an element of $\mathscr{B}_{\rho^*}$. The next lemma sharpens this result and gives a probabilistic interpretation of f.

LEMMA A1.3. *Assume Conditions* A1.1, A1.2, *and* A1.3. *Let f be determined by the equations* ($i \geqslant 0$)

$$f(i+1) - f(i) = (p_{i,i+1})^{-1} \left[\sum_{j=0}^{i-1} p_{ij}(f(i) - f(j)) + p_{i0} f(0) \right]$$

and $f(0) = 1$. *Then* $(\rho^*)^n / f(n) \to 0$ *as* $n \to \infty$.

PROOF. Let $q^n(i)$ denote the probability that, starting in state $i < n$, state n is visited before the next visit to state 0. The recurrence equations for $q^n(i)$ are:

$$q^n(i) = \sum_{j=1}^{i+1} p_{ij} q^n(j), \qquad i \in [0, n-1],$$

with $q^n(n) = 1$. It is clear from (7) that $q^n(i) = f(i)/f(n)$ is a solution to these equations. In particular $1 = f(0) = q^n(0)f(n)$, so that $1/f(n) = q^n(0) =$ the probability that, starting in state 0, state n is reached before the next visit to state 0.

To prove that $(\rho^*)^n/f(n) \to 0$, consider the function $s(\cdot)$ on S defined by $s(n) := (\rho^*)^n$. Since $s \in \mathscr{B}_{\rho^*}$, the sum $\sum_{t=0}^{\infty} P_1^t s$ is finite. Considering the interpretation of f, it is clear that $\sum_{t=0}^{\infty}(P_1^t s)(0) \geqslant \sum_{n=0}^{\infty}(\rho^*)^n/f(n)$, and hence $(\rho^*)^n/f(n) \to 0$, as $n \to \infty$.

∎

The results derived so far can be used to calculate $v(i)$, $i \geqslant 0$, in an efficient way. For Lemma A1.3 and the fact that $v \in \mathscr{B}_{\rho^*}$ (from Theorem A1.2) imply that $\lim_{n\to\infty} v(n)/f(n) = 0$ and therefore $v(0) = \lim_{n\to\infty} -w(n)/f(n)$. Since both $f(n)$ and $w(n)$ can be calculated recursively from (10), this formula can be used to approximate $v(0)$. Once $v(0)$ is known, the values of $v(1)$, $v(2)$, etc., can also be calculated recursively, using (10). See Wijngaard [9] for details.

Our main use of the results from this appendix, however, will be to distinguish between w and v in the case where $r = r_\alpha - g\tau_\alpha$: the case of g-revised reward in a semi-Markov decision process under strategy α. Since $v = w + fv(0)$, it follows from Theorem A1.2 and Lemma A1.3 that $w \in \mathscr{B}_{\rho^*}$ *if and only if* $v(0) = 0$, in which case $w = v$. We use this fact in §2 (Theorem 6) to provide a check on whether $v^g(0) = 0$ and thus on whether $g = g^*$.

Appendix 2. Proofs of theorems in §1.

PROOF OF THEOREM 1. For any strategy α, let Q_α denote the transition operator of the imbedded Markov chain on $[0, n_0 - 1]$ generated by α, where n_0 is defined in Condition 6. That is,

$$(Q_\alpha f)(i) := \sum_{t=0}^{\infty} \left(P_{\alpha n_0}^t P_\alpha f\right)(i), \qquad i \geqslant 0,$$

for all functions f with support in $[0, n_0 - 1]$ (cf. Appendix 1). Let $Q_{\alpha 1}$ be the restriction of Q_α to $[1, n_0 - 1]$, so that $Q_{\alpha 1} = \sum_{t=0}^{\infty} P_{\alpha n_0}^t P_{\alpha 1}$.

Let α be such that $g_\alpha(i) \geqslant g_{\hat{\alpha}}$ for all i. First we prove that the imbedded Markov chain on $[0, n_0 - 1]$ generated by α is nondegenerate. Define $l_{\alpha_0^n}(i) := \lim_{n\to\infty}(P_{\alpha n_0}^t \mathbf{1})(i) = \Pr\{$system stays in $[n_0, \infty)$ forever under strategy $\alpha |$ start in state $i\}$. For $i \geqslant n_0$, we have $g \leqslant g_\alpha(i) \leqslant l_{\alpha n_0}(i) g_0 + (1 - l_{\alpha n_0}(i)) M_2/L$ and hence, since $g_0 < g \leqslant M_2/L$,

$$l_{\alpha n_0}(i) \leqslant [M_2/L - g_{\hat{\alpha}}]/[M_2/L - g_0] := \delta < 1.$$

It follows that $l_{\alpha n_0}(i) \leqslant \delta < 1$ for all $i \geqslant 0$. But

$$l_{\alpha n_0} = P_{\alpha n_0} l_{\alpha n_0} = \lim P_{\alpha n_0}^t l_{\alpha n_0} \leqslant l_{\alpha n_0},$$

which implies that $l_{\alpha n_0}(i) = 0$ for all i. This means that under strategy α for each starting state i the set $[0, n_0 - 1]$ is reached with probability one. In other words, the imbedded chain on $[0, n_0 - 1]$ is nondegenerate, as claimed.

Now we modify the process under strategy α in the following way. In the first place we set $r_\alpha(i)$ equal to M_2 for $i < n_0$ and equal to $g_0 \tau_\alpha(i)$ for $i \geqslant n_0$. In the second place, if the state remains outside the set $[0, n_0)$ for more than N transitions, we only count the contribution of the first N transitions, in computing both the expected total reward and the expected total time until the nth transition. The result is a semi-Markov process with rewards, defined on the state space $\{(i, n): i = 0, 1, \ldots;\ n = 0, 1, \ldots,$

and $n = 0$ for $i < n_0\}$. Here the component n records the cumulative number of transitions since the last visit to $[0, n_0)$ and both reward and transition time for states (i, n) with $n \geqslant N$ are equal to 0.

Let $g_\alpha^N(i, 0)$ be the average reward for this process, defined in the usual way. It follows from the definition of n_0 and the hypothesis that $g_\alpha(i) \geqslant g_{\hat{\alpha}}$ that $g_\alpha^N(i, 0) \geqslant g_{\hat{\alpha}}$ for all i. Let $R_\alpha^N(i)$ and $T_\alpha^N(i)$ denote the expected reward and expected time for the modified process until the next visit to a state $(j, 0)$ with $j < n_0$, starting in $(i, 0)$ with $i < n_0$. Both expectations are finite. From the way in which rewards are defined for the modified process, it follows that

$$R_\alpha^N(i) = M_2 + g_0\big(T_\alpha^N(i) - \tau_\alpha(i)\big), \qquad i < n_0. \tag{11}$$

(In particular, for $i < n_0 - 1$ the skip-free transition structure implies that $T_\alpha^N(i) = \tau_\alpha(i)$ and $R_\alpha^n(i) = M_2$.)

Condition 4 implies that the expected reward and the expected time until the next visit to state $(0, 0)$, starting in state (i, n), are finite for all i and n. For starting state $(i, 0)$ with $i < n_0$ these expectations equal, respectively, $\sum_{t=0}^{\infty}(Q_{\alpha 1}^t R_\alpha^N)(i)$ and $\sum_{t=0}^{\infty}(Q_{\alpha 1}^t T_\alpha^N)(i)$. Hence we can write the average reward $g_\alpha^N(i, 0)$ in the following way (for all i)

$$g_\alpha^N(i, 0) = \left[\sum_{t=0}^{\infty}(Q_{\alpha 1}^t R_\alpha^N)(0)\right] \Big/ \left[\sum_{t=0}^{\infty}(Q_{\alpha 1}^t T_\alpha^N)(0). \tag{12}\right.$$

Using (11), (12), and the fact that $g_\alpha^N(i, 0) \geqslant g_{\hat{\alpha}}$, we obtain

$$\begin{aligned} g_{\hat{\alpha}} \sum_{t=0}^{\infty}(Q_{\alpha 1}^t T_\alpha^N)(0) \leqslant M_2 \sum_{t=0}^{\infty}(Q_{\alpha 1}^t \mathbf{1})(0) + g_0 \sum_{t=0}^{\infty}(Q_{\alpha 1}^t T_\alpha^N)(0) \\ - g_0 \sum_{t=0}^{\infty}(Q_{\alpha 1}^t \tau_\alpha)(0), \end{aligned}$$

so that

$$(g_{\hat{\alpha}} - g_0) \sum_{t=0}^{\infty}(Q_{\alpha 1}^t T_\alpha^N)(0) \leqslant M_2 \sum_{t=0}^{\infty}(Q_{\alpha 1}^t \mathbf{1})(0). \tag{13}$$

The summation on the right-hand side of (13) equals the expected number of visits to $[0, n_0 - 1]$ before the next visit to 0, starting from state 0, which is finite by Condition 4. Thus (13) implies that $\sum_{t=0}^{\infty}(Q_{\alpha 1}^t T_\alpha^N)(0)$ has an upper bound that is independent of N. Let $T_\alpha(i)$ denote the expected time in the original process until the next visit to $[0, n_0 - 1]$, starting in i. Then

$$\sum_{t=0}^{\infty}(Q_{\alpha 1}^t T_\alpha^N)(0) \to \sum_{t=0}^{\infty}(Q_{\alpha 1}^t T_\alpha)(0),$$

as $N \to \infty$. The latter sum is therefore finite, and is equal to the sum $\sum_{t=0}^{\infty}(P_{\alpha 1}^t \tau_\alpha)(0)$. Since $p_{i,i+1}(\alpha(i)) > 0$ for all i, it follows that $\sum_{t=0}^{\infty}(P_{\alpha 1}^t \tau_\alpha)(i)$ is finite for all i. We may therefore write

$$g_\alpha(i) = g_\alpha := \left[\sum_{t=0}^{\infty}(P_{\alpha 1}^t r_\alpha)(0)\right] \Big/ \left[\sum_{t=0}^{\infty}(P_{\alpha 1}^t \tau_\alpha)(0),\right.$$

for all $i \geqslant 0$. Since $g_\alpha \geqslant g_{\hat{\alpha}}$, the sum $\sum_{t=0}^{\infty}(P_{\alpha 1}^t r_\alpha)(0)$ is also finite and because $p_{i,i+1}(\alpha(i)) > 0$, we may conclude that the sum $\sum_{t=0}^{\infty}(P_{\alpha 1}^t r_\alpha)(i)$ is finite for all i. This completes the proof of Theorem 1. ∎

Proof of Lemma 2. By Conditions 3 and 4 the function v^g is bounded above. Let b be an upper bound. The finiteness of $\sum_{t=0}^{\infty}(P_{\alpha 1}^n \tau_\alpha)(i)$ implies that $\lim_{n\to\infty}(P_{\alpha 1}^n b)(i) = 0$.

Let $v_\alpha^g(i)$ be the total expected accumulation of $r_\alpha - g\tau_\alpha$ until the next visit to 0, starting in i and following strategy α. So $v_\alpha^g(i) = \sum_{t=0}^\infty P_{\alpha 1}^t(r_\alpha - g\tau_\alpha)(i)$. The finiteness of $\sum_{t=0}^\infty (P_{\alpha 1}^t r_\alpha)(i)$ and $\sum_{t=0}^\infty (P_{\alpha 1}^t \tau_\alpha)(i)$ implies that $v_\alpha^g(i)$ is finite and that $\lim_{n\to\infty}(P_{\alpha 1}^n v_\alpha^g)(i) = 0$. The desired result then follows from the fact that $v_\alpha^g(i) \leqslant v^g(i) \leqslant b$ and $\lim_{n\to\infty}(P_{\alpha 1}^n b)(i) = 0$ for all i. ∎

Proof of Lemma 3. Let $g \geqslant g_0$ be arbitrary. For all i,a let

$$f^l(i,a) := \left[r(i,a) - g\tau(i,a)\right]\mathbf{1}(i < n_0) \quad \text{and} \quad f^h(i,a) := \left[r(i,a) - g\tau(i,a)\right]\mathbf{1}(i \geqslant n_0),$$

where n_0 is from Condition 6. Conditions 3 and 4 imply that the total accumulation of f^l until the next visit to 0, starting in i, is bounded over all strategies. Let α be a strategy that attains the maximum in (4), so that $v_\alpha^g(i) = v^g(i)$. Then the total accumulation of f^h until the next visit to 0, starting in i and following α, is finite. This implies, by the definition of f^h, that the expected total accumulation of both r and τ until the next visit to 0, starting in i and following α, is finite. Let $T(i)$ denote the expected total accumulation of τ until the next visit to 0, starting in i and following α. Let $\epsilon > 0$ be arbitrary. Choose $\delta \leqslant \epsilon / T_\alpha(i)$. The total accumulation of $r - (g + \delta)\tau$ until the next visit to 0, starting in i and following α, is equal to $v^g(i) - \delta T_\alpha(i)$, and hence $v^{g+\delta}(i) \geqslant v^g - \delta T_\alpha(i) \geqslant v^g(i) - \epsilon$. By the definition of v^g it is clear that $v^{g+\delta}(i) \leqslant v^g(i)$. We have thus established continuity of $v^g(i)$ in $g \geqslant g_0$. ∎

Proof of Theorem 4. Let $P := P_{\alpha_{g^*}}$, $r := r_{\alpha_{g^*}}$, $\tau := \tau_{\alpha_{g^*}}$. It follows from equation (4) that

$$v^{g^*} = \sum_{t=0}^{n-1} P_1^t(r - g^*\tau) + P_1^n v^{g^*}.$$

By the arguments used in the proof of Lemma 3, the expected total accumulation of both r and τ until the next visit to 0, starting in i and following α_{g^*}, is finite. Therefore, by Lemma 2, $\lim_{n\to\infty}(P_1^n v^{g^*})(i) = 0$ for all i. Hence

$$v^{g^*}(0) = \lim_{n\to\infty} \sum_{t=0}^{n-1} P_1^t(r - g^*)(0) = \sum_{t=0}^{\infty} (P_1^t r)(0) - g^* \sum_{t=0}^{\infty} (P_1^t \tau)(0).$$

Since $v^{g^*}(0) = 0$ this implies that the average reward under α_{g^*} equals

$$g^* = \left[\sum_{t=0}^{\infty} (P_1^t r)(0)\right] \Big/ \left[\sum_{t=0}^{\infty} (P_1^t \tau)(0)\right].$$

Now suppose there is another strategy α with average reward $g_\alpha \geqslant g^*$. By Lemmas 1 and 2, $\lim_{n\to\infty}(P_{\alpha 1}^n v^{g^*})(i) = 0$ and

$$g_\alpha(i) = \left[\sum_{t=0}^{\infty} (P_1^t r_\alpha)(0)\right] \Big/ \left[\sum_{t=0}^{\infty} (P_1^t \tau_\alpha)(0)\right] := g_\alpha.$$

Moreover,

$$\begin{aligned} v^{g^*}(i) &\geqslant r_\alpha(i) - g^*\tau_\alpha(i) + \left(P_{\alpha 1} v^{g^*}\right)(i) \geqslant \cdots \\ &\geqslant \sum_{t=0}^{n-1} (P_{\alpha 1}^t r)(i) - g^* \sum_{t=0}^{n-1} (P_{\alpha 1}^t \tau_\alpha)(i) + \left(P_{\alpha 1}^n v^{g^*}\right)(i). \end{aligned}$$

Hence $0 = v^{g^*}(0) \geqslant (g_\alpha - g^*)\sum_{t=0}^\infty (P_{\alpha 1}^t \tau_\alpha)(0)$, so that $g_\alpha \leqslant g^*$. ∎

References

[1] Derman, C. (1970). *Finite-Stage Markovian Decision Processes.* Academic Press, New York.

[2] Keilson, J. (1962). The Use of Green's Functions in the Study of Random Walks, with Applications to Queueing Theory. *J. Math. and Physics* **41** 42–52.

[3] ———. (1965). *Green's Function Methods in Probability Theory*. Griffin, London.

[4] Low, D. (1974). Optimal Pricing Policies for an M/M/s Queue. *Oper. Res.* **22** 545–561.

[5] ———. (1974). Optimal Pricing for an Unbounded Queue. *IBM J. Res. and Develop.* **18** 290–302.

[6] Miller, B. (1981). Countable-State Average-Cost Regenerative Stopping Problems. *J. Appl. Probab.* **18** 361–377.

[7] Ross, S. (1970). *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco.

[8] Schäl, M. (1975). Conditions for Optimality in Dynamic Programming and for the Limit of n-Stage Optimal Policies to be Optimal. *Z. Wahrsch. Verw. Gebiete* **32** 179–196.

[9] Wijngaard, J. (1978). A Direct Numerical Method for a Class of Queueing Problems. *Management Sci.* **24** 1441–1447.

WIJNGAARD: DEPARTMENT OF INDUSTRIAL ENGINEERING, EINDHOVEN UNIVERSITY OF TECHNOLOGY, EINDHOVEN, THE NETHERLANDS

STIDHAM: DEPARTMENT OF INDUSTRIAL ENGINEERING, NORTH CAROLINA STATE UNIVERSITY, BOX 7906, RALEIGH, NORTH CAROLINA 27695-7906