JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 141, 303–317 (1989)

Solving Infinite Horizon Discounted Markov Decision Process Problems for a Range of Discount Factors

D. J. WHITE

Department of Systems Engineering, University of Virginia, Charlottesville, Virginia 22901

Submitted by E. Stanley Lee

Received May 18, 1987

1. INTRODUCTION

In this paper we will assume the following framework. There is a finite state set I, with $i \in I$ as its generic member, $1 \leqslant i \leqslant m$. For each $i \in I$, there is a finite action set $K(i)$, with $k \in K(i)$ as its generic member. For each $i \in I$, $k \in K(i)$, there is a transition probability, $p(i, j, k)$, that if at a decision epoch the state is $i \in I$, and if action $k \in K(i)$ is taken, then the state will be $j \in I$ at the next decision epoch, and there is an immediate reward, $r(i, k)$, with $0 \leqslant r(i, k) \leqslant M < \infty$; there is a discount factor τ in the interval $[0, \rho]$, for some fixed $\rho < 1$. We will be interested in values of τ within this range, and we will parameterize τ by a parameter $t \in [0, 1]$, so that $\tau = t\rho$. The actual values of t to be studied will depend upon the questions we may wish to answer and how these might efficiently be answered.

A pure Markov decision rule is a function $\delta: I \to K = \bigcup_{i \in I} K(i)$, where if $i \in I$ then $\delta(i) \in K(i)$. We will be concerned with maximizing the infinite horizon discounted rewards, and we do not need to consider either time dependent decision rules, or those which are a function of the complete history of the process up to a specified decision epoch or decision rules which select an action with a specified probability (see van der Wal [4]). A general policy π is an infinite sequence of decision rules which determine an action at each decision epoch, as a function of the past history, with some probability. In the light of the previous remark we need only consider policies of the form $\pi = (\delta)^\infty$, where $\delta \in \Delta$, the set of pure Markov decision rules, and $(\delta)^\infty$ simply means the application of δ an infinite number of times.

If $v_t(i)$ is the maximum infinite horizon expected discounted reward, with

0022-247X/89 $3.00

discount factor $\tau = t\rho$, beginning in state $i \in I$, then $v_t: I \to R$ is a unique solution to the following functional optimality equation (see White [7]),

$$v_t = T_t v_t,$$

where

$$[T_t v_t](i) = \max_{k \in K(i)} [T_t^k v_t](i)$$

$$[T_t^k y](i) = r(i, k) + t\rho \sum_{j \in I} p(i, j, k)\, y(j)$$

for all $y: I \to R$.

We write the above optimality equation in the following form

$$v_t = T_t v_t = \max_{\delta \in \Delta} [T_t^\delta v_t] = \max_{\delta \in \Delta} [r^\delta + t\rho P^\delta v_t].$$

In addition, all optimal pure Markov policies π are of the form $\pi = (\delta)^\infty$ where

$$\delta(i) \in \arg\max_{\delta \in \Delta} [[T_t^\delta v_t](i)],$$

where, in the case of multiple arg max solutions any arg max solution may be chosen.

The motivation for this paper comes, to some extent, from the paper by Smallwood [3]. In that paper a method is given for finding optimal policies for the whole range of τ values in the set $[0, 1]$. Broadly speaking that paper proceeds by finding critical τ values where there is a change in optimal policy for some state $i \in I$. Policies will be optimal for a set of (not necessarily adjacent) closed intervals of $[0, 1]$, and Smallwood's procedure generates these closed intervals.

In this paper we will adopt a different, exploratory approach which relates v_t to $v_{t+\sigma}$, where σ may take any value in $[0, 1-t)$, but where the computational schemes may be more profitably used if σ is small.

We emphasize the exploratory nature of the paper since, although algorithms are specified, and their properties examined, it is by no means conclusive that the procedures suggested will turn out to be efficient. Nonetheless it does present ideas which might profitably be explored in more detail at a later stage.

In the next section we will present various algorithms and their convergence properties. There are two objectives in mind, viz,

(a) to find approximations for v_t over the range $[0, 1]$;

(b) to find approximations for $v_{t+\sigma}$ when v_t is given.

In the analysis σ may take any value in the range $[0, 1-t)$ (for $t+\sigma$) and we may also consider all values of $t\sigma$ for $0 \leqslant t \leqslant 1/\sigma$. When σ is small then we will get better approximation results.

Algorithms A1, A2 will be aimed at objective (a), and algorithms B1–B3 will be aimed at objective (b).

2. Algorithms

First of all we present an elementary result which will be required for some of the future analysis. Throughout this paper $\|x\|$, for $x \in R^m$, will be the maximum norm.

R.1. $0 \leqslant \sigma \leqslant 1-t$: $0 \leqslant v_{t+\sigma} - v_t \leqslant (\sigma\rho/(t-(t+\sigma)\rho))\,\|v_t\|$.

Proof. Let δ_t arg $\max_{\delta \in \Delta}[T^{\delta} v_t]$. Then

$$\begin{aligned} v_{t+\sigma} - v_t &= T_{t+\sigma}^{\delta_{t+\sigma}} v_{t+\sigma} - T_t^{\delta_t} v_t \\ &\geqslant T_{t+\sigma}^{\delta_t} v_{t+\sigma} - T_t^{\delta_t} v_t \\ &= \sigma\rho P^{\delta_t} v_{t+\sigma} + t\rho P^{\delta_t}(v_{t+\sigma} - v_t). \end{aligned}$$

Hence

$$v_{t+\sigma} - v_t \geqslant \sigma\rho(I_m - t\rho P^{\delta_t})^{-1} P^{\delta_t} v_{t+\sigma} \geqslant 0$$

when I_m is the mth order identity matrix. Also

$$\begin{aligned} v_{t+\sigma} = v_t &\leqslant T_{t+\sigma}^{\delta_{t+\sigma}} v_{t+\sigma} - T_t^{\delta_{t+\sigma}} v_t \\ &= \sigma\rho P^{\delta_{t+\sigma}} v_t + (t+\sigma)\rho P_{t+\sigma}^{\delta_{t+\sigma}}(v_{t+\sigma} - v_t). \end{aligned}$$

Hence

$$v_{t+\sigma} - v_t \leqslant \sigma\rho(I_m - (t+\sigma)\rho P^{\delta_{t+\sigma}})^{-1} P^{\delta_{t+\sigma}} v_t.$$

The requisite result now follows. ∎

Let us now consider our algorithms, noting that Algorithms A. 1 and A.2 are aimed at objective (a) of Section 1.

Algorithm A.1

(i) **Set** $t = 0$, $w_0 = v_0$;

(ii) **Stop** if $t = 1$, and otherwise **go to** (iii);

(iii) **Find**

$$\tau_{t+\sigma} \in \arg\max_{\delta \in \Delta} [T^{\delta}_{t+\sigma} w_t];$$

(iv) **Solve** the following equation for $w_{t+\sigma}$:

$$w_{t+\sigma} = T^{\tau_{t+\sigma}}_{t+\sigma} w_{t+\sigma};$$

(v) **go to** (ii).

We then have the following result.

R.A.1. (i) $0 \leqslant v_{t+\sigma} - w_{t+\sigma} \leqslant (t+\sigma)\,\rho\,\|(v_t - w_t)\| + ((t+\sigma)/(1-(t+\sigma)\,\rho))\,\|v_t\|$, *for* $0 \leqslant t \leqslant 1-\sigma$;

(ii) $0 \leqslant v - w \leqslant (\rho^2 M/(1-\rho)^3)(1-\rho^s)/s$ *where* $v = v_1$, $w = w_1$, $s = 1/\sigma$;

(iii) $0 \leqslant v_{t+\sigma} - w_{t+\sigma} \leqslant (\rho(t+\sigma)/(1-\rho(t+\sigma)))\,\|w_{t+\sigma} - w_t\|$ *for* $0 \leqslant t \leqslant 1-\sigma$.

Proof. For $0 \leqslant \sigma \leqslant 1-t$ we have the following:

$$\begin{aligned} w_{t+\sigma} - w_t &= T^{\tau_{t+\sigma}}_{t+\sigma} w_{t+\sigma} - T^{\tau_t}_t w_t \\ &= (T^{\tau_{t+\sigma}}_{t+\sigma} w_{t+\sigma} - T^{\tau_{t+\sigma}}_{t+\sigma} w_t) \\ &\quad + (T^{\tau_{t+\sigma}}_{t+\sigma} w_t - T^{\tau_t}_{t+\sigma} w_t) \\ &\quad + (T^{\tau_t}_{t+\sigma} w_t - T^{\tau_t}_t w_t \\ &= A + B + C, \text{ say.} \end{aligned}$$

By definition

$$\begin{aligned} B &= T^{\tau_{t+\sigma}}_{t+\sigma} w_t - T^{\tau_t}_{t+\sigma} w_t \\ &= T_{t+\sigma} w_t - T^{\tau_t}_{t+\sigma} w_t \geqslant 0. \end{aligned}$$

Clearly we have

$$C = \sigma\rho P^t w_t \geqslant 0.$$

Hence

$$\begin{aligned} w_{t+\sigma} - w_t &\geqslant T^{\tau_{t+\sigma}}_{t+\sigma} w_{t+\sigma} - T^{\tau_{t+\sigma}}_{t+\sigma} w_t. \\ &= (t+\sigma)\,\rho P^{\tau_{t+\sigma}}(w_{t+\sigma} - w_t). \end{aligned}$$

Hence

$$(I_m - (t+\sigma)\,\rho P^{\tau_{t+\sigma}})(w_{t+\sigma} - w_t) \geqslant 0.$$

Hence

$$w_{t+\sigma} \geqslant w_t.$$

Let us now consider $v_{t+\sigma} - w_{t+\sigma}$. We have

$$\begin{aligned} v_{t+\sigma} - w_{t+\sigma} &= T_{t+\sigma} v_{t+\sigma} - T_{t+\sigma}^{\tau_{t+\sigma}} w_{t+\sigma} \\ &= (T_{t+\sigma} v_t - T_{t+\sigma} w_t) \\ &\quad + (T_{t+\sigma} v_{t+\sigma} - T_{t+\sigma} v_t) \\ &\quad + (T_{t+\sigma}^{\tau_{t+\sigma}} w_t - T_{t+\sigma}^{\tau_{t+\sigma}} w_{t+\sigma}) \\ &= A + B + C, \text{ say,} \end{aligned}$$

after noting that $T_{t+\sigma} w_t = T_{t+\sigma}^{\tau_{t+\sigma}} w_t$ by definition.

Since $w_{t+\sigma} \geqslant w_t$, we see that $C \leqslant 0$. Then using result R.1 we obtain the following, after noting that, clearly, $v_{t+\sigma} \geqslant w_{t+\sigma}$:

$$v_{t+\sigma} - w_{t+\sigma} \leqslant ((t+\sigma)\rho)(\max_{\delta \in \Delta}[P^\delta(v_t - W_t)] + \max_{\delta \in \Delta}[P^\delta(v_{t+\sigma} - v_t)].$$

The requisite result (i) follows from this and result R.1.

For part (ii) of the result, we note that, since $0 \leqslant t + \sigma \leqslant 1$, we have the following:

$$0 \leqslant v_{t+\sigma} - w_{t+\sigma} \leqslant \rho(v_t - w_t) + \sigma\rho^2 M/(1-\rho)^2.$$

Then, if $t = s\sigma$, we obtain the following:

$$0 \leqslant v_t - w_t \leqslant (\rho^2 M/(1-\rho)^3)(1-\rho^s)/s.$$

This is the requisite result (ii).

For result (iii) we slightly change the approach in (i). We have the following:

$$\begin{aligned} 0 \leqslant v_{t+\sigma} - w_{t+\sigma} &= T_{t+\sigma} v_{t+\sigma} - T_{t+\sigma} w_{t+\sigma} \\ &\leqslant T_{t+\sigma} v_{t+\sigma} - T_{t+\sigma} w_t \qquad (\text{since } w_{t+\sigma} \geqslant w_t) \\ &= T_{t+\sigma} v_{t+\sigma} - T_{t+\sigma} w_{t+\sigma} + T_{t+\sigma} v_{t+\sigma} - T_{t+\sigma} w_t \\ &\leqslant (t+\sigma)\rho(\max_{\delta \in \Delta}[P^\delta(v_{t+\sigma} - w_{t+\sigma})] + \max_{\delta \in \Delta}[P^\delta(w_{t+\sigma} - w_t)]. \end{aligned}$$

Hence

$$\begin{aligned} 0 &\leqslant v_{t+\sigma} - w_{t+\sigma} \\ &\leqslant \rho(t+\sigma) \max_{\delta \in \Delta}[(I_m - (t+\sigma)\rho P^\delta)^{-1} \max_{\tau \in \Delta}[P^\tau(w_{t+\sigma} - w_t)]]. \end{aligned}$$

This gives the requisite result (iii). ∎

It is to be noted that the error bound in (ii) is a "prior" bound, calculated in advance of the actual results, whereas (iii) is a "posterior" bound, calculated as the actual results are obtained.

Algorithm A.1. has a "policy space" component in step (iv). The next algorithm is a form of the "successive approximations" algorithm and contains no policy space iteration.

ALGORITHM A.2

(i) **Set** $t=0$, and $u_0=v_0$;

(ii) **stop** if $t=1$, and otherwise **go to** (iii);

(iii) **find**

$$u_{t+\sigma}=T_{t+\sigma}u_t;$$

(iv) **go to** (ii).

We then have the following result.

R.A.2. (i) $0\leqslant v_{t+\sigma}-u_{t+\sigma}\leqslant((t+\sigma)\,\rho/(1-(t+\sigma)\,\rho))\,\|u_{t+\sigma}-u_t\|$ *for* $0\leqslant t\leqslant 1-\sigma$;

(ii) $0\leqslant v-u\leqslant(s!(\rho/s)^{s+1}/(1-\rho))(\sum_{z=0}^{s-1}(s-z)(\rho/s)^{-z}/z!)\,M$, *where* $v=v_1$, $u=u_1$, $s=1/\sigma$.

Proof. For $t+\sigma\leqslant 1$ we have the following. Let $\mu_{t+\sigma}\in \arg\max_{\delta\in\Delta}[T^{\delta}_{t+\sigma}u_t]$. Then

$$\begin{aligned}v_{t+\sigma}-u_{t+\sigma}&\geqslant(t+\sigma)\,\rho P^{\mu_{t+\sigma}}(v_{t+\sigma}-u_t)\\&\geqslant(t+\sigma)\,\rho P^{\mu_{t+\sigma}}(v_t-u_t)\end{aligned}$$

since, clearly, $v_{t+\sigma}\geqslant v_t$. Repeating the argument, and since $v_0=u_0$, we see that

$$v_t\geqslant u_t \qquad \text{for all } t.$$

We now obtain the reverse inequality as follows. Let

$$\delta_{t+\sigma}\in\arg\max_{\delta\in\Delta}[T^{\delta}_{t+\sigma}v_{t+\sigma}].$$

Then

$$\begin{aligned}v_{t+\sigma}-u_{t+\sigma}&\leqslant(t+\sigma)\,\rho P^{\delta_{t+\sigma}}(v_{t+\sigma}-u_t)\\&=(t+\sigma)\,\rho P^{\delta_{t+\sigma}}(v_{t+\sigma}-u_{t+\sigma})\\&\quad+(t+\sigma)\,\rho P^{\delta_{t+\sigma}}(u_{t+\sigma}-u_t).\end{aligned}$$

Hence

$$v_{t+\sigma} - u_{t+\sigma} \leqslant (t+\sigma)\,\rho\,\|u_{t+\sigma} - u_t\|/(1-(t+\sigma)\,\rho).$$

Thus result (i) is established.

Now

$$\begin{aligned} u_{t+\sigma} - u_t &\geqslant (t+\sigma)\,\rho P^{\mu_t} u_t - t\rho P^{\mu_t} u_{t+\sigma} \\ &\geqslant \sigma\rho P^{\mu_t}(u_t - u_{t-\sigma}). \end{aligned}$$

Repeating the process, since $\mu_\sigma \geqslant \mu_0$, for all t we see that $u_{t+\sigma} \geqslant u_t$. We may now obtain a reverse inquality as follows:

$$\begin{aligned} u_{t+\sigma} - u_t &\leqslant (t+\sigma)\,\rho P^{\mu_{t+\sigma}} u_t - t\rho P^{\mu_t} u_{t-\sigma} \\ &= t\rho P^{\mu_{t-\sigma}}(u_t - u_{t-\sigma}) + \sigma\rho P^{\mu_{t+\sigma}} u_t. \end{aligned}$$

Hence

$$\|u_{t+\sigma} - u_t\| \leqslant t\rho\,\|u_t - u_{t-\sigma}\| + \sigma\rho\,\|u_t\|.$$

Putting $t = s\sigma$, we have the following:

$$\|u_{t+\sigma} - u_t\| \leqslant s!\,(\sigma\rho)^{s+1} \sum_{k=0}^{s} \frac{(\sigma\rho)^{-k}}{k!}\,\|u_{k\sigma}\|.$$

Now

$$\|u_{k\sigma}\| \leqslant \|u_0\| + k\sigma\rho\,\|u_{(k-1)\sigma}\|.$$

Repeating this process we obtain the following:

$$\|u_{k\sigma}\| \leqslant k!\left(\sum_{r=0}^{k} \frac{(\sigma\rho)^r}{(k-r)!}\right)\|u_0\|.$$

Hence

$$\|u_{t+\sigma} - u_t\| \leqslant s!\,(\sigma\rho)^{s+1}\left(\sum_{k=0}^{s}\sum_{r=0}^{k} \frac{(\sigma\rho)^{r-k}}{(k-r)!}\right)\|u_0\|.$$

Hence, with $t = s\sigma$, we obtain the following:

$$\|v_{(s+1)\sigma} - u_{(s+1)\sigma}\| \leqslant \frac{((s+1)!\,(\sigma\rho)^{s+2}}{(1-(s+1)\,\sigma\rho)}\left(\sum_{k=0}^{s}\sum_{r=0}^{k} \frac{(\sigma\rho)^{r-k}}{(k-r)!}\right)\|u_0\|.$$

Thus

$$\|v_t - u_t\| \leqslant \frac{s!\,(\sigma\rho)^{s+1}}{(1-s\sigma\rho)} \left(\sum_{k=0}^{s-1} \sum_{r=0}^{k} \frac{(\sigma\rho)^{r-k}}{(k-r)!} \right) \|u_0\|$$

$$\leqslant \frac{s!\,(\sigma\rho)^{s+1}}{(1-s\sigma\rho)} \left(\sum_{q=1}^{s} \frac{q(\sigma\rho)^{q-s}}{(s-q)!} \right) \|u_0\|$$

$$= \frac{s!(\sigma\rho)^{s+1}}{(1-s\sigma\rho)} \left(\sum_{z=0}^{s-1} \frac{(s-z)(\sigma\rho)^{-z}}{z!} \right) \|u_0\|.$$

This gives the requisite result (ii) when $\sigma = 1/s$. ∎

The bound given in (ii) is quite complicated, and it is also difficult to see how it varies as s increases (σ decreases). At the very least, it should tend to zero as s tends to infinity.

From the analysis, an upper bound is as follows:

$$0 \leqslant v - u \leqslant ((t+\sigma)\,\rho/(1-(t+\sigma)\,\rho)\,s!\,(\rho/s)^{s+1} \left(\sum_{k=0}^{s} (\rho/s)^{-k}/k! \right) \|v\|.$$

We may use Stirling's formula (see Feller [1, p. 50]),

$$s! = (1+\varepsilon(s))(2\pi)^{1/2}\, s^{s+1/2} e^{-s},$$

where $\varepsilon(s)$ tends to zero as s tends to infinity. Then we have the following:

$$0 \leqslant v - u \leqslant (2\pi)^{1/2}\,((t+\sigma)\,\rho^2/(1-(t+\sigma)\,\rho))\,\|v\|\, s^{-1/2} e^{-s} e^{s}(1+\varepsilon(s))$$

$$= (2\pi)^{1/2}\,((t+\sigma)\,\rho^2/(1-(t+\sigma)\,\rho)(1-\rho))\,M(1+\varepsilon(s))\, s^{-1/2}.$$

The right hand side of this inequality tends to zero as s tends to ∞. For $s \geqslant 10$ (see Feller [1, p. 50]), $|\varepsilon(s)| \leqslant 8 \times 10^{-3}$.

The error bound in (i) is a "posterior" bound and that in (ii) is a "prior" bound.

It is natural to examine whether or not u_t is greater than or equal to v_t. The following result holds.

R.A.1/2. $0 \leqslant t \leqslant 1$: $w_t \geqslant u_t$.

Proof. We have the following for $0 \leqslant t \leqslant 1-\sigma$,

$$u_{t+\sigma} = T_{t+\sigma} u_t$$

$$w_{t+\sigma} = T_{t+\sigma}^{T_{t+\sigma}} w_{t+\sigma},$$

where $\tau_{t+\sigma} \in \arg\max_{\delta \in \Delta} [T^{\delta} w_t]$. Then

$$\begin{aligned} w_{t+\sigma} - u_{t+\sigma} &= T_{t+\sigma}^{\tau_{t+\sigma}} w_{t+\sigma} - T_{t+\sigma} u_t \\ &\geqslant T_{t+\sigma}^{\tau_{t+\sigma}} w_t - T_{t+\sigma} u_t \\ &= T_{t+\sigma} w_t - T_{t+\sigma} u_t \\ &\geqslant 0 \text{ if } w_t \geqslant u_t. \end{aligned}$$

Since $w_0 = u_0$, the requisite result follows. ∎

We now deal with algorithms for objective (b) in Section 1. The first one is the standard successive approximations algorithm (e.g., see White [7]).

ALGORITHM B.1

(i) **let** $n = 0$ and $v_{0,t+\sigma} = v_t$;

(ii) **if** $n = N$, **stop**, and otherwise **go to** (iii);

(iii) **find**

$$v_{n+1,t+\sigma} = T_{t+\sigma} v_{n,t+\sigma}.$$

We then have the following result.

R.B.1. $n \geqslant 1$:

(i) $0 \leqslant v_{t+\sigma} - v_{n,t+\sigma} \leqslant ((t+\sigma)\rho/(1-(t+\sigma)\rho))\, \|v_{n,t+\sigma} - v_{n-1,t+\sigma}\|$;

(ii) $0 \leqslant v_{t+\sigma} - v_{n,t+\sigma} \leqslant ((t+\sigma)\rho)^n 0\rho/(1-(t+\sigma)\rho))\, \|v_t\|$.

Proof. We have the following. $1 \leqslant n \leqslant N$. Let $\gamma_{t+\sigma} \in \arg\max_{\delta \in \Delta} [T_{t+\sigma}^{\delta} v_{n,t+\sigma}]$. Then

$$\begin{aligned} v_{t+\sigma} - v_{n,t+\sigma} &\geqslant T_{t+\sigma}^{\gamma_{t+\sigma}} v_{t+\sigma} - T^{\gamma_{t+\sigma}} v_{n-1,t+\sigma} \\ &= (t+\sigma)\rho P^{\gamma_{t+\sigma}} (v_{t+\sigma} - v_{n-1,t+\sigma}). \end{aligned}$$

Now

$$\begin{aligned} v_{1,t+\sigma} - v_{0,t+\sigma} &= T_{t+\sigma} v_t - v_t \\ &\geqslant T_t v_t - v_t = 0. \end{aligned}$$

Hence

$$v_{t+\sigma} \geqslant v_{n,t+\sigma}.$$

Similarly we have the following. Let $\delta_{t+\sigma} \in \arg\max_{\delta \in \Delta} [T^{\delta}_{t+\sigma} v_{t+\sigma}]$.

Then

$$v_{t+\sigma} - v_{n,t+\sigma} \leqslant T^{\delta_{t+\sigma}}_{t+\sigma} v_{t+\sigma} - T^{\delta_{t+\sigma}}_{t+\sigma} v_{n-1,t+\sigma}$$
$$= (t+\sigma)\,\rho\,(P^{\delta_{t+\sigma}}(v_{t+\sigma} - v_{n,t+\sigma}) + P^{\delta_{t+\sigma}}(v_{n,t+\sigma} - v_{n-1,t+\sigma})).$$

Hence

$$0 \leqslant v_{t+\sigma} - v_{n,t+\sigma} \leqslant (t+\sigma)\,\rho(I_m - (t+\sigma)\,\rho P^{\delta_{t+\sigma}})^{-1}\, P^{\delta_{t+\sigma}}(v_{n,t+\sigma} - v_{n-1,t+\sigma}).$$

This gives result (i).

Using a similar analysis we have the following:

$$0 \leqslant v_{t+\sigma} - v_{n,t+\sigma} \leqslant (t+\sigma)\,\rho P^{\delta_{t+\sigma}}(v_{t+\sigma} - v_{n-1,t+\sigma}).$$

Also

$$v_{1,t+\sigma} - v_{0,t+\sigma} = T_{t+\sigma} v_t - v_t$$
$$\leqslant \sigma\rho P^{\delta_{t+\sigma}} v_t.$$

Result (ii) then follows. ∎

Error bound (i) is a "posterior" bound and error bound (ii) is a "prior" bound.

The next two algorithms contain a policy space step which is similar to, but not identical with, that in White *et al.* [5]. There is also a similarity to the successive over relaxation method, which is policy space oriented (see White [9]).

ALGORITHM B.2

(i) **Set** $n = 0$ and $v^0_{t+\sigma} = v_t$;

(ii) **if** $n = N$ **stop**, and otherwise **go to** (iii);

(iii) **solve** the following equation for $v^{n+1}_{t+\sigma}$:

$$v^{n+1}_{t+\sigma} = T_{t,n} v^{n+1}_{t+\sigma}$$

where
for $y: I \to R^m$

$$[T_{t,n}\, y](i) = \max_{k \in K(i)} [r(i,k) + \sigma\rho v^n_{t+\sigma} + t\rho y]$$

which may be written as

$$T_{t,n}\, y = \max_{\delta \in \Delta} [r^{\delta} + \sigma\rho v^{n}_{t+\sigma} + tpy];$$

(iv) **go to** (ii).

We then have the following result.

R.B.2. $n \geqslant 1$.

(i) $0 \leqslant v_{t+\sigma} - v^{n}_{t+\sigma} \leqslant (\sigma(t+\sigma)\,\rho^2/(1-(t+\sigma)\,\rho))\,\|v^{n}_{t+\sigma} - v^{n-1}_{t+\sigma}\|$;

(ii) $0 \leqslant v_{t+\sigma} - v^{n}_{t+\sigma} \leqslant (\sigma\rho/(1-t\rho))^{n}\,(\sigma\rho/(1-(t+\sigma)\,\rho))\,\|v_t\|$.

Proof. Let $\phi_{t+\sigma} \in \arg\max_{\delta \in \Delta} [T^{\delta}_{t,n-1} v^{n}_{t+\sigma}]$. Then

$$\begin{aligned} v_{t+\sigma} - v^{n}_{t+\sigma} &\geqslant T^{\phi_{t+\sigma}}_{t+\sigma} v_{t+\sigma} - T^{\phi_{t+\sigma}}_{t,n-1} v^{n}_{t+\sigma} \\ &= \sigma\rho P^{\phi_{t+\sigma}}(v_{t+\sigma} - v^{n-1}_{t+\sigma}) + t\rho P^{\phi_{t+\sigma}}(v_{t+\sigma} - v^{n}_{t+\sigma}). \end{aligned}$$

Hence

$$v_{t+\sigma} - v^{n}_{t+\sigma} \geqslant \sigma\rho(I_m - t\rho P^{\phi_{t+\sigma}})^{-1}\, P^{\phi_{t+\sigma}}(v_{t+\sigma} - v^{n-1}_{t+\sigma}).$$

Now

$$v_{t+\sigma} - v^{0}_{t+\sigma} = v_{t+\sigma} - v_t \geqslant 0.$$

Hence, repeating the above procedure we obtain

$$v_{t+\sigma} \geqslant v^{n}_{t+\sigma}.$$

With a similar analysis we may obtain reverse inequalities as follows:

$$\begin{aligned} v_{t+\sigma} - v^{n}_{t+\sigma} &\leqslant T^{\delta_{t+\sigma}}_{t+\sigma} v_{t+\sigma} - T^{\delta_{t+\sigma}}_{t,n-1} v^{n}_{t+\sigma} \\ &= \sigma\rho P^{\delta_{t+\sigma}}(v_{t+\sigma} - v^{n-1}_{t+\sigma}) + t\rho P^{\delta_{t+\sigma}}(v_{t+\sigma} - v^{n}_{t+\sigma}). \end{aligned}$$

Hence

$$0 \leqslant v_{t+\sigma} - v^{n}_{t+\sigma} \leqslant \sigma\rho(I_m - t\rho P^{\delta_{t+\sigma}})^{-1}\, P^{\delta_{t+\sigma}}(v_{t+\sigma} - v^{n}_{t+\sigma}).$$

Repeating this analysis we obtain the following:

$$\begin{aligned} 0 \leqslant v_{t+\sigma} - v^{n}_{t+\sigma} &\leqslant (\sigma\rho/(1-t\rho))^{n}\,\|v_{t+\sigma} - v_t\| \\ &\leqslant (\sigma\rho/(1-t\rho))^{n}\,(\sigma\rho/(1-(t+\sigma)\,\rho))\,\|v_t\|. \end{aligned}$$

This is result (ii).

We similarly have the following:

$$\begin{aligned} 0 \leqslant v_{t+\sigma} - v^{v}_{t+\sigma} &\leqslant T^{\delta_{t+\sigma}}_{t+\sigma} v_{t+\sigma} - T^{\delta_{t+\sigma}}_{t,n-1} v^{n}_{t+\sigma} \\ &= T^{\delta_{t+\sigma}}_{t+\sigma} v_{t+\sigma} - T^{\delta_{t+\sigma}}_{t+\sigma} v^{n}_{t+\sigma} \\ &\quad + T^{\delta_{t+\sigma}}_{t+\sigma} v^{n}_{t+\sigma} - T^{\delta_{t+\sigma}}_{t,n-1} v^{n}_{t+\sigma}. \end{aligned}$$

Hence

$$0 \leqslant v_{t+\sigma} - v^{n}_{t+\sigma} \leqslant (t+\sigma)\,\rho (I_m - (t+\sigma)\,\rho P^{\delta_{t+\sigma}})^{-1}\,(\sigma \rho P^{\delta_{t+\sigma}}(v^{n}_{t+\sigma} - v^{n-1}_{t+\sigma})).$$

Repeating this analysis gives results (i). ∎

The error bounds given in (i) are "posterior" bounds and those in (ii) are "prior" bounds.

Algorithm B.3. reverses the roles of t and σ and the results are similar.

Algorithm B.3

As for Algorithm B.2, but replacing $T^{\delta}_{t,n}$ by $\tilde{T}^{\delta}_{t,n}$ where

$$\tilde{T}_{t,n}\, y = \max_{\delta \in \Delta} [r^{\delta} + t\rho P^{\delta} \tilde{v}^{n}_{t+\sigma} + \sigma \rho P^{\delta} y].$$

The sequence $\{v^{n}_{t}\}$ will be replaced by $\{\tilde{v}^{n}_{t}\}$.

The following result is obtained by an automatic application of result R.B.2.

R.B.3. $n \geqslant 1$:

(i) $0 \leqslant v_{t+\sigma} - \tilde{v}^{n}_{t+\sigma} \leqslant (t(t+\rho)\,\rho^2/(1-(t+\sigma)\,\rho))\,\|\tilde{v}^{n}_{t+\sigma} - \tilde{v}^{n-1}_{t+\sigma}\|$;

(ii) $0 \leqslant v_{t+\sigma} - \tilde{v}^{n}_{t+\sigma} \leqslant (t\rho/(1-\sigma\rho))^{n}\,(\sigma\rho/(1-(t+\sigma)\,\rho))\,\|v_t\|$.

The significance of the difference between Algorithms B.2 and B.3 lies in the use of these when σ is small with respect to t and the convergence rates are different. However, the policy space phases (iii) also have different effective discount factors. We will return to these later on.

Finally let us derive some results comparing $\{v_{n,t+\sigma}\}$, $\{v^{n}_{t+\sigma}\}$, and $\{\tilde{v}^{n}_{t+\sigma}\}$. These are as follows.

R.B.1/2/3. $1 \leqslant n \leqslant N$:

(i) $v^{n}_{t+\sigma} \geqslant v_{n,t+\sigma}$;

(ii) $\tilde{v}^{n}_{t+\sigma} \geqslant v_{n,t+\sigma}$;

(iii) $v^{n}_{t+\sigma} \geqslant \tilde{v}^{n}_{t+\sigma}$ *if and only if* $t \geqslant \sigma$.

Proof. (i) $1 \leqslant n \leqslant N-1$:

$$\begin{aligned} v_{t+\sigma}^{n+1} - v_{n+1,t+\sigma} &= T_{t,n} v_{t+\sigma}^{n+1} - T_{t+\sigma} v_{n,t+\sigma} \\ &\geqslant T_{t,n} v_{t+\sigma}^{n} - T_{t+\sigma} v_{n,t+\sigma} \\ &\geqslant (t+\sigma)\, \rho \min_{\delta \in \Delta} [P^{\delta}(v_{t+\sigma}^{n} - v_{n,t+\sigma})]. \end{aligned}$$

Now

$$v_{t+\sigma}^{0} = v_{0,t+\sigma} = v_t.$$

Hence the requisite result (i) follows by induction.

(ii) This follows in the same way as for (i).

(iii) We see that the following is true.

$$v_{t+\sigma}^{n} = \max_{\delta \in \Delta} [(I_m - \sigma\rho P^{\delta})^{-1} (r^{\delta} + \sigma\rho v_{t+\sigma}^{n-1})]$$

$$\tilde{v}_{t+\sigma}^{n} = \max_{\delta \in \Delta} [(I_m - \sigma\rho P^{\delta})^{-1} (r^{\delta} + t\rho \tilde{v}_{t+\sigma}^{n-1})].$$

Now $v_{t+\sigma}^{0} = v_{t+\sigma}^{0} = v_t$. Let $t \geqslant \sigma$. Assume that $v_{t+\sigma}^{n-1} \geqslant \tilde{v}_{t+\sigma}^{n-1}$. Then

$$\begin{aligned} v_{t+\sigma}^{n} - \tilde{v}_{t+\sigma}^{n} &\geqslant \min_{\delta \in \Delta} [(\sigma\rho(I_m - t\rho P^{\delta})^{-1} - t\rho(I_m - \sigma\rho P^{\delta})^{-1})\, \tilde{v}_{t+\sigma}^{n-1}] \\ &= \min_{\delta \in \Delta} \left[\sum_{s=0}^{\infty} (\sigma\rho(t\rho)^{s} - t\rho(\sigma\rho)^{s})(P^{\delta})^{s}\, \tilde{v}_{t+\sigma}^{n-1} \right] \\ &= \min_{\delta \in \Delta} \left[\sigma t \sum_{s=0}^{\infty} \rho^{s+1}(t^{s-1} - \sigma^{s-1})(P^{\delta})^{s}\, v_{t+\sigma}^{n-1} \right] \\ &\geqslant 0. \end{aligned}$$

Hence the requisite result (iii) follows, noting that if $t \leqslant \sigma$ a similar analysis will apply. ∎

3. Discussion of Algorithms

It should be emphasized that this is an exploratory paper whose purpose is to examine the properties of various algorithms for solving infinite horizon discounted Markov decision process problems for specified sets of discount factors. A more detailed consideration of the computational complication will be necessary before the effectiveness of any can be established.

The same theme governing all the algorithms is that of being able to use computational information from the approximate solution to a problem

with discount factor $t\rho$ to facilitate the computation of an approximate solution to a problem with discount factors $(t+\sigma)\rho$. In some cases σ may be small with respect to t, and one would expect good approximations. In others σ may not be small with respect to t, but the separation of $(t+\sigma)\rho$ into $t\rho$ and $\sigma\rho$ produces two smaller discount factors, which may lead to faster convergence rates of the algorithms.

Algorithms A1 and A2 give schemes for solving problems over a whole interval $[0, \rho]$ of discount factors, with ρ being specified in advance. The schemes are clearly extendable to a more general range $[\underline{\rho}, \bar{\rho}]$ if required.

Algorithm A1 might be of some use for actually solving a problem with discount factor ρ by a series of successive approximations, but with the succession being along the discount factor scale. The standard successive approximation algorithm (see White [7]) has an error bound proportional to ρ^s (for s iterations); whereas the Algorithm A1 has an error bound proportional to $1/s$. For some cases A1 will be better, but this is likely to be for small s values. It is to be noted that the bound in (ii) of result R.A.1 is a crude bound, and a better bound is obtainable by a more detailed evaluation of results R.A.1(i).

Result R.A.1(iii) allows us to stop calculations at any stage when the current value of $\|w_{t+\sigma} - w_t\|$ is small enough.

For Algorithm A1 there is the "policy space" step (iii). This requires more time than a standard successive approximation step, which may make the algorithm unattractive for solving a specific problem with discount factor ρ.

Algorithm A2 is clearly of little use to solve a specific problem with discount factor ρ, since the standard successive approximation procedure, for the same number of iterations, will always produce higher value functions. For the purpose of approximating the solution to all problems with discount factors in the range $[0, \rho]$, for the same number of iterations, as for Algorithm A1, each iteration is easier, but this may be offset by the fact that Algorithm A1 produces a dominating sequence of value functions (see Result R.A.1/2). Also, as explained after the proof of Result R.A.2, for large s values the approximate asymptotic bound calculated for Algorithm A2 is inversely proportional to $s^{1/2}$, and not to s as is the case with Algorithm A1.

Result R.A.2(i) allows us to stop calculations at any stage when the current values of $\|u_{t+\sigma} - u_t\|$ are small enough.

Algorithm B1–B3 are aimed at approximating $v_{t+\sigma}$, given v_t. The convergence rates for these algorithms are, respectively, $(t+\sigma)\rho$, $\sigma\rho/(1-t\rho)$, $t\rho/(1-\sigma\rho)$. If σ is small enough then Algorithm B2 has the faster convergence rates, and its value function sequence dominates those of Algorithms B1 and B3 (see Result R.B.1/2/3(i), (iii)). However, the policy space step (iii) can be time consuming, and some attention to finding

suitable approximation schemes for the step is required. The effective discount factor ($\sigma\rho$) for Algorithm B3 at the policy space step is, for small σ, lower than that for Algorithm B2 ($t\rho$) and this may lead to easier approximation schemes, based, perhaps, on successive approximations.

Even when σ is not small with respect to t there may be some advantage in using Algorithms B2 or B3. For example, if $\sigma = t = \frac{1}{2}$, then the convergence rates of the algorithms (at each iteration) B1–B3 became respectively ρ, $\frac{1}{2}\rho/(1-\frac{1}{2}\rho)$, $\frac{1}{2}\rho/(1-\frac{1}{2}\rho)$, and the latter two are always faster. The policy space steps of Algorithms B2 and B3 still pose problems, but again their effective discount factors (for the policy space step) are $t\rho$ and $\sigma\rho$, respectively.

Finally, the sequences $\{u_t\}$ (Algorithm A2), $\{v_{n,t}\}$ (Algorithm B1) do not, except perhaps accidentally, correspond to the value functions for a policy. However, these may be used to generate approximate value functions which do; e.g., see Porteus [2] and White [6, 8], where, in the case of White [6], errors in discount factors are studied.

REFERENCES

1. W. FELLER, "An Introduction to Probability Theory and Its Application," Wiley, New York, 1957.
2. E. L. PORTEUS, Some bounds for discounted sequential decision processes, *Man. Sci.* **18** (1971), 7–11.
3. R. D. SMALLWOOD, Optimum policy regions for Markov processes with discounting, *Oper. Res.* **21** (1966), 1071–1088.
4. J. VAN DER WAL, "Stochastic Dynamic Programming," Mathematisch Centrum, Amsterdam, 1981.
5. C. C. WHITE, L. C. THOMAS, W. T. SCHERER, Reward revision for discounted Markov decision problems, *Oper. Res.* **33** 1299–1315.
6. D. J. WHITE, Infinite horizon Markov decision processes with unknown or variable discount factors, *European J. Oper. Res.* **28** (1987), 96–100.
7. D. J. WHITE, "Finite Dynamic Programming," Wiley, New York, 1978.
8. D. J. WHITE, The determination of approximately optimal policies in Markov decision processes by the use of bounds, *J. ORS* **33** (1982), 253–257.
9. D. J. WHITE, A survey of algorithms for some restricted classes of Markov decision problems, *Proc. Oper. Res.* **8** (1979), 103–121.