A stopping time-based policy iteration algorithm for average reward Markov decision processes

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

A stopping time-based policy iteration algorithm for average reward Markov decision processes

by

J. van der Wal

Memorandum COSOR 78-11

Eindhoven, May 1978

The Netherlands

A stopping time-based policy iteration algorithm for average reward Markov decision processes

by

J. van der Wal

Abstract

We consider Howard's policy iteration algorithm for multichained finite state and action Markov decision processes at the criterion of average reward per unit time. Using stopping times as has been done by Wessels in the total reward case we obtain a set of policy improvement steps, among which Gauss Seidel, which as we show give convergent algorithms and produce average optimal strategies.

1. Introduction and notations

In this paper we deal with the finite state and action Markov decision process (MDP) at the criterion of average reward per unit time. We will consider Howard's policy iteration method [6] and we introduce stopping times, as has been done for total reward MDP by Wessels [11] and Van Nunen and Wessels [9] to obtain a set of policy improvement steps. Among them the Gauss-Seidel step suggested by Hastings [3]. And we show that each of these stopping time-based algorithms terminates with an average optimal strategy.

So we are looking at a discrete time MDP with finite state space $S := \{1,2,\ldots,N\}$ and finite action space A. If in state i action a is taken the immediate reward is $r(i,a)$ and a transition is made to state j with probability $p(j|i,a)$. A strategy π in this MDP is any sequence $(\pi_0,\pi_1,\ldots)$ of mappings π_n from $(S \times A)^n \times S$ [the set of histories upto time n] into A. [The restriction to nonrandomized strategies is not relevant]. Each $i \in S$ and π determine a probability $\mathbb{P}_{i,\pi}$ on $(S \times A)^\infty$ and a stochastic process $\{(X_n,A_n),\ n = 0,1,\ldots\}$ where X_n is the state and A_n the action at time n. The expectation with respect to $\mathbb{P}_{i,\pi}$ is denoted by $\mathbb{E}_{i,\pi}$ and $\mathbb{E}_\pi(\cdot)$ denotes the N-vector with i-th component $\mathbb{E}_{i,\pi}(\cdot)$.

A strategy for which there exists a map $f : S \to A$ such that $\pi_n(h_n,i) = f(i)$ for all n, all $h_n \in (S \times A)^n$ and $i \in S$ will be called a stationary strategy or a policy and we denote it by f. By r_f we denote the N-vector with i-th component $r(i,f(i))$ and similarly P_f denotes the matrix with $P_f(i,j) = p(j|i,f(i))$. And we define

$$P_f^* := \lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} P_f^k .$$

Let f be a policy, $g_f \in \mathbb{R}^N$ denote the gain or average reward per unit time and $v_f \in \mathbb{R}^N$ be the bias term [$v_f = \lim_{n\to\infty} \Sigma_{k=0}^{n-1} P_f^k r_f - n g_f$ if the Markov chain corresponding to f is aperiodic]. Then (g_f,v_f) is

the unique solution of

$$(1,1;f)\quad \begin{cases} r_f + P_f v = v + g \\ P_f g = g \\ P_f^* v = 0 \ . \end{cases}$$

The standard policy improvement step may now be formulated as follows:

Let f be the actual policy then find an improved policy h as follows.
Let D(i,f) and E(i,f), $i \in S$ be defined by

$$D(i,f) := \{a \in A \mid \Sigma_j p(j|i,a) g_f(j) = \max_k \Sigma_j p(j|i,k) g_f(j)\}$$

$$E(i,f) := \{a \in D(i,f) \mid r(i,a) + \Sigma_j p(j|i,a) v_f(j) = \max_{k \in D(i,f)} \{r(i,k) + \Sigma_j p(j|i,k) v_f(j)\} \ .$$

Take h to be any policy with $h(i) \in E(i,f)$, such that if $f(i) \in E(i,f)$ then $h(i) = f(i)$, $i \in S$.

The gain g_h of a policy h obtained in this way is at least equal to the gain g_f of f and if $g_h = g_f$ on S then the bias term v_h of h is at least equal to the bias term of f : $v_h \geq v_f$. Now one may perform the policy improvement step again on h, etc. until a policy is found that cannot be improved anymore, h^* say. Then h^* is optimal gain, i.e. $g_{h^*} \geq g_f$ for all policies f.
This algorithm is known as Howard's policy iteration algorithm [6]. Actually his formulation was slightly different. In Howard's version the equation $P_f^* v = 0$ in (1,1;f) is replaced by the condition that in each irreducible class of P_f one component of v is set equal to zero. The formulation given here seems to stem from Blackwell [1]. A convergence proof can be found in Derman [2].
In this paper we propose a different policy improvement step which we will formulate by means of stopping times. In all generality a stopping time is a function τ on S^∞ with the property

$$\tau(i_0,i_1,\ldots,i_n,i_{n+1},i_{n+2},\ldots) = n \Rightarrow \tau(i_0,\ldots,i_n,j_{n+1},j_{n+2},\ldots) = n$$

for all $j_{n+1},j_{n+2},\ldots \in S$.
For reasons that will become clear in the sequel [we will come back to it in section 7] we restrict ourselves to a special class of stopping times: the set of nonzero, finite and transition memoryless stopping times. By nonzero we mean:

$\tau(i_0,i_1,\ldots) \geq 1$ for all $i_0,i_1,\ldots \in S$, by finite we mean $\mathbb{P}_{i,\pi}(\tau < \infty) = 1$ for all i,π. So whether a stopping time is finite or not may depend on the MDP under consideration. The term transition memoryless refers to the property that stopping only occurs after a transition from i to j for special pairs (i,j). Formally, there exists a subset $T \subset S^2$ such that $\tau(i_0,i_1,\ldots) = n \Leftrightarrow (i_k,i_{k+1}) \notin T$, $k = 0,\ldots,n-2$, $(i_{n-1},i_n) \in T$.
[So transition memoryless stopping times are automatically nonzero.]
In the sequel all stopping times will be assumed to be nonzero, finite and transition memoryless.
One may show that these stopping times τ are also exponentially bounded [i.e. for all π there exists an M and $\alpha < 1$ such that $\mathbb{P}_{i,\pi}(\tau > n) < M\alpha^n$, $i \in S$]
So, for any stopping time τ and strategy π, we may define the vector $r_{\tau,\pi}$ and the matrices $P_{\tau,\pi}$ and $Q_{\tau,\pi}$ by

$$r_{\tau,\pi}(i) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\tau-1} r(X_n,A_n) , \qquad i \in S$$

$$P_{\tau,\pi}(i,j) := \mathbb{P}_{i,\pi}(X_\tau = j) , \qquad i,j \in S$$

$$Q_{\tau,\pi}(i,j) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\tau-1} \delta(X_n,j) , \quad i,j \in S$$

where $\delta(k,j) = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases}$

Now we are able to give a rough formulation of the modified policy improvement step we propose.

> Let f be any policy and (g_f,v_f) solve $(1,1;f)$.
> Then, first maximize $P_{\tau,\cdot}g_f$ and secondly use the remaining freedom to maximize $r_{\tau,\cdot} + P_{\tau,\cdot}v_f - Q_{\tau,\cdot}\bar{g}$, where $\bar{g} = \max_\pi P_{\tau,\pi}g_f$.

We subtract the term $Q_{\tau,\cdot}g$ since we must compare $r_{\tau,\cdot}$ with $Q_{\tau,\cdot}$ times the average amount we expect to get, which is at least $\bar{g}$ [for a strategy prescribing a maximizer $\bar{\pi}$ of $P_{\tau,\cdot}g_f$ upto time τ and f thereafter].
We will see that the [one of the] maximizer[s] in the modified policy improvement step is a policy.

Notice that the stopping time characterized by the set $T = \{(i,j) \mid j \geq i\}$ corresponds to the Gauss-Seidel policy improvement step.

In section 2 we will motivate the modified policy improvement step by considering the discounted MDP when the discountfactor tends to 1.
In section 3 we give the full description of the modified improvement step. Section 4 gives some preliminary results needed to show in section 5 that the modified improvement step produces a better policy.
Section 6 shows that repeated application of the modified improvement step yields an average optimal strategy.

Before we proceed with section 2 we introduce one more notation which will simplify our formulas there.
Let f be any policy then we may split up the matrix P_f into the matrices $\bar{P}_f$ and $\tilde{P}_f$ defined by

$$\bar{P}_f(i,j) := \begin{cases} P_f(i,j) & \text{if } j \notin T_i \\ 0 & \text{else} \end{cases}$$

where $T_i := \{j \in S \mid (i,j) \in T\}$

$$\tilde{P}_f(i,j) := \begin{cases} P_f(i,j) & \text{if } j \in T_i \\ 0 & \text{else} \end{cases}$$

Then we have for stationary strategies

Lemma 1.1 Let f be a policy then

(i) $r_{\tau,f} = (I - \bar{P}_f)^{-1} r_f$

(ii) $P_{\tau,f} = (I - \bar{P}_f)^{-1} \tilde{P}_f$

(iii) $Q_{\tau,f} = (I - \bar{P}_f)^{-1}$.

Proof. The proof is straightforward. For example

$$r_{\tau,f} = r_f + \bar{P}_f r_f + \bar{P}_f^2 r_f + \ldots = (I - \bar{P}_f)^{-1} r_f \, . \qquad \square$$

2. Motivation of the modified policy improvement step

In this section we give a motivation for the improvement step by considering the discounted MDP when the discountfactor tends to 1.
For a policy f we have the following equation (cf. Blackwell [1])

$$(2.1) \qquad v_{\beta,f} = (1-\beta)^{-1} g_f + v_f + o(1) \qquad (\beta \uparrow 1) ,$$

where $v_{\beta,f}$ denotes the total expected discounted return under f:

$$v_{\beta,f} = \mathbb{E}_f \sum_{n=0}^{\infty} \beta^n r(X_n, A_n) .$$

And also the functional equation

$$v_{\beta,f} = r_{\beta,\tau,f} + P_{\beta,\tau,f} v_{\beta,f} ,$$

with

$$r_{\beta,\tau,f} = \mathbb{E}_f \sum_{n=0}^{\tau-1} \beta^n r(X_n, A_n) ,$$

the expected discounted reward upto time τ, and

$$P_{\beta,\tau,f}(i,j) = \sum_{n=1}^{\infty} \beta^n \mathbb{P}_{i,f}(X_\tau = j, \tau = n) .$$

The following discounted analogon of lemma 1.1 is straightforward

Lemma 2.1.

$$r_{\beta,\tau,f} = (I - \beta \bar{P}_f)^{-1} r \quad \text{and} \quad P_{\beta,\tau,f} = (I - \beta \bar{P}_f)^{-1} \beta \tilde{P} .$$

If we apply a successive approximation step on $v_{\beta,f}$ then we maximize (restricting ourselves to policies, which is allowed by theorem 3.2 in Wessels [11])

$$(2.2) \qquad r_{\beta,\tau,h} + P_{\beta,\tau,h}[(1-\beta)^{-1} g_f + v_f + o(1)] \qquad (\beta \uparrow 1) ,$$

or

$$(2.3) \qquad (I - \beta \bar{P}_h)^{-1} [r_h + [\tilde{P} - (1-\beta)\tilde{P}][(1-\beta)^{-1} g_f + v_f + o(1)]] \qquad (\beta \uparrow 1) .$$

We need the following lemma

Lemma 2.2.

$$(I - \beta\bar{P}_h)^{-1} = \sum_{n=0}^{\infty} (-1)^n (1-\beta)^n \{\bar{P}_h (I - \bar{P}_h)^{-1}\}^n (I - \bar{P}_h)^{-1} .$$

Proof. $(I - \beta\bar{P}_h)^{-1} = (I - \bar{P}_h + (1-\beta)\bar{P}_h)^{-1}$.

Further $\bar{P}_h = \bar{P}_h(I - \bar{P}_h)(I - \bar{P}_h)^{-1} = (I - \bar{P}_h)\bar{P}_h(I - \bar{P}_h)^{-1}$.

So $(I - \beta\bar{P}_h)^{-1} = [I - \bar{P}_h + (1-\beta)(I - \bar{P}_h)\bar{P}_h(I - \bar{P}_h)^{-1}]^{-1}$

$$= [I + (1-\beta)\bar{P}_h(I - \bar{P}_h)^{-1}]^{-1}(I - \bar{P}_h)^{-1} .$$

Expanding the first term on the rhs now yields the desired result. □

Substituting the result of lemma 2.2 in (2.3) we get

$$(2.4) \quad (1-\beta)^{-1}(I - \bar{P}_h)^{-1}\tilde{P}g_f + (I - \bar{P}_h)^{-1}r_h - (I - \bar{P}_h)^{-1}\tilde{P}g_f +$$

$$(I - \bar{P}_h)^{-1}\tilde{P}v_f - \bar{P}_h(I - \bar{P}_h)^{-1}(I - \bar{P}_h)^{-1}\tilde{P}g_f + o(1) \quad (\beta \uparrow 1) .$$

Which, using lemma 1.1 and taking together the constant terms with $\tilde{P}_h g_f$, reduces to

$$(2.5) \quad (1-\beta)^{-1}P_{\tau,h}g_f + r_{\tau,h} + P_{\tau,h}v_f - Q_{\tau,h}P_{\tau,h}g_f + o(1) \quad (\beta \uparrow 1) .$$

So, if we want to maximize (2.2) for β sufficiently close to 1, our first concern will be to maximize $P_{\tau,h}g_f$. Once we have done that we will maximize $r_{\tau,h} + P_{\tau,h}v_f - Q_{\tau,h}P_{\tau,h}g_f$. And this is precisely the improvement step we proposed in section 1.

On the other hand if policy f itself is optimal in both tests then we have for all h

$$(2.6) \quad r_{\beta,\tau,h} + P_{\beta,\tau,h}v_{\beta,f} \leq v_{\beta,f} + o(1) \qquad (\beta \uparrow 1)$$

If we iterate this and use $P^n_{\beta,\tau,h} e \le \beta^n e$ $(e = (1,1,\ldots,1)^T)$[as follows from $\tau \ge 1$]then we get

$$(2.7) \qquad \sum_{n=0}^{N-1} P^n_{\beta,\tau,h} r_{\beta,\tau,h} + P^N_{\beta,\tau,h} v_{\beta,f} \le v_{\beta,f} + (1 + \beta + \ldots + \beta^{N-1}) o(1) .$$

Letting N tend to infinity gives

$$(2.8) \qquad v_{\beta,h} \le v_{\beta,f} + o((1-\beta)^{-1}) \qquad (\beta \uparrow 1) .$$

So f must be an optimal gain policy.

The modified policy improvement step

First we will give the full discription of the policy improvement step we already introduced roughly.

Let f be a policy and (g_f, v_f) solve $(1,1;f)$.
Define $\psi \in \mathbb{R}^N$ by

$$(3.1) \qquad \max_h P_{\tau,h} g_f = g_f + \psi$$

and for all $i \in S$ the set $A(i,f)$ of all $a \in A$ for which

$$(3.2) \qquad \sum_{j \in T_i} p(j|i,a) g_f(j) + \sum_{j \notin T_i} p(j|i,a)(g_f + \psi)(j) = (g_f + \psi)(i) .$$

Let $H(f)$ be the set of policies which only use actions from $A(i,f)$, $i \in S$ $[h \in H(f) \Leftrightarrow h(i) \in A(i,f),\ i \in S]$.
We will see that all policies from $H(f)$ maximize (3.1).
Define γ by

$$(3.3) \qquad \max_{h \in H(f)} \{r_{\tau,h} + P_{\tau,h} v_f - Q_{\tau,h}(g_f + \psi)\} = v_f + \gamma$$

and for all $i \in S$ the set $B(i,f)$ as the set of $a \in A(i,f)$ with

$$(3.4) \qquad r(i,a) - g_f(i) - \psi(i) + \sum_{j \in T_i} p(j|i,a) v_f(j) + \sum_{j \notin T_i} p(j|i,a)(v_f + \gamma)(j) = (v_f + \gamma)(i) .$$

Define an improved policy h with $h(i) \in B(i,f)$ and if $f(i) \in B(i,f)$ then $h(i) = f(i)$.

In section 5 we will show that policy h is indeed an improvement of f. But before we come to that we want to consider the modified improvement step in more detail.
In the maximization (3.1) we only consider stationary strategies. The following lemma shows that nothing can be gained by considering other strategies.

Lemma 3.1. For all $w \in \mathbb{R}^N$

$$\sup_{\pi} P_{\tau,\pi} w = \max_{h} P_{\tau,h} w \ . \tag{3.5}$$

Proof. We follow the line of proof of theorem 3.2 in Wessels [11]. First define the following MDP

$$\bar{r}(i,a) := \sum_{j \in T_i} p(j|i,a) w(j)$$

$$\bar{p}(j|i,a) := \begin{cases} 0 & \text{if } j \in T_i \\ p(j|i,a) & \text{if } j \notin T_i \end{cases} \ . \tag{3.6}$$

One easily sees that this newly defined MDP is equivalent to the original problem. From the fact that we demanded τ to be finite one may obtain that the new process is N-stage contracting, thus contracting (cf. section 7 in Van Hee and Wessels [4]. Hence, for example by theorem 3.1.(ii) in Van Nunen and Wessels [10], we can restrict ourselves to policies. Since both MDP are equivalent we can restrict ourselves to policies in the original maximization problem as well. □

If we write out the functional equation of the MDP (3.6) with $w = g_f$ then we get

$$\max_{a} \{\bar{r}(i,a) + \sum_{j} \bar{p}(j|i,a)(g_f + \psi)(j)\} = (g_f + \psi)(i) \ , \qquad i \in S \ . \tag{3.7}$$

We see that the set A(i,f) is precisely the set of actions which attain the maximum in (3.7). And the lhs of (3.2) is at most equal to the rhs.

So A(i,f) is the set of conserving actions (cf. Hordijk [5]). Since τ is finite all strategies are equalizing, so any strategy consisting of actions from A(i,f) only must maximize (3.1). Reversively one may show that strategies maximizing (3.1) effectively use only actions from the sets A(i,f). Therefore, our fist aim being to maximize (3,1), it is natural to consider only strategies taking actions from A(i,f) in the maximization (3.3). Moreover let $\Pi(f)$ be the reduced set of strategies that only use actions from A(i,f) then we have

Lemma 3.2. For all $v,w \in \mathbb{R}^N$

$$\max_{\pi\in\Pi(f)} \{r_{\tau,\pi} + P_{\tau,\pi}v - Q_{\tau,\pi}w\} = \max_{h\in H(f)} \{r_{\tau,h} + P_{\tau,h}v - Q_{\tau,h}w\} . \tag{3.8}$$

Proof. The proof is almost identical to the proof of lemma 3.1 with only the action sets being smaller and a different reward structure

$$\bar{\bar{r}}(i,a) := r(i,a) - w(i) + \sum_{j\in T_i} p(j|i,a)v(j) , \qquad i \in S,\ a \in A(i,f) \tag{3.9}$$

and $$\bar{\bar{p}}(j|i,a) := \bar{p}(j|i,a) , \qquad i \in S,\ a \in A(i,f) .$$ □

As a result we can restrict ourselves in the maximization (3.7) to policies from H(f).

Now the functional equation of MDP (3.9) with $v = v_f$ and $w = g_f + \psi$ becomes

$$\max_{a\in A(i,f)} \{\bar{\bar{r}}(i,a) + \sum_j \bar{\bar{p}}(j|i,a)(v_f + \gamma)(j)\} = (v_f + \gamma)(i) , \quad i \in S . \tag{3.10}$$

So B(i,f) is the set of actions attaining the maximum in (3.10). And the sets B(i,f) are the sets of conserving actions which according to the same reasoning as before produce all policies maximizing (3.3).

4. Some preliminary results

In the next section we will need a result about the chainstructure of the matrices $P_{\tau,h}$.

Lemma 4.1.

(i) Each irreducible class of P_h contains exactly one irreducible class of $P_{\tau,h}$ and possibly some states which are transient under P_τ.

(ii) If i belongs to an irreducible class C of P_h then $P_{\tau,h}(i,j) > 0$ only if j belongs to the irreducible class of $P_{\tau,h}$ contained in C.

Proof. The proof is not difficult and follows from the fact that τ is transition memoryless. We prefer to omit it there. □

For more general stopping times lemma 4.1 need not hold. For example, let

$$P_h = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

and $\tau \equiv 3$, then P_h has only one irreducible class : $\{1,2,3\}$, but $P_{\tau,h}$ three: $\{1\}$, $\{2\}$ and $\{3\}$.

An important consequence of lemma 4.1(ii) which we will use in section 5 is

Corollary 4.2. If w is constant on an irreducible class C of $P_{\tau,h}$ then $P_{\tau,h}w$ is constant on the class of P_h containing C.

Another result we will need is formulated in the following lemma.

Lemma 4.3. For the solution (g_f,v_f) of (1,1;f) we have

(4.1) (i) $P_{\tau,f}g_f = g_f$

(4.2) (ii) $r_{\tau,f} + P_{\tau,f}v_f - Q_{\tau,f}g_f = v_f$.

Proof. We only prove (i), the proof of (ii) being similar.
From lemma 1.1(ii), the definition of $\bar{P}_f$ and $\tilde{P}_f$ and $P_f g_f = g_f$ we have

$$P_{\tau,f}g_f = (I - \bar{P}_f)^{-1}\tilde{P}_f g_f = (I - \bar{P}_f)^{-1}(P_f - \bar{P}_f)g_f$$

$$= (I - \bar{P}_f)^{-1}(I - \bar{P}_f)g_f = g_f \ .$$

□

5. The modified policy improvement step improves

In this section we show that the modified policy improvement step from section 3 yields an improved policy h. I.e., let f be a policy and (g_f, v_f) solve (1,1;f) and let h be a policy obtained from f by the modified improvement step and (g_h, v_h) solve (1,1;h). Then either

(i) $g_h \geq g_f$ and $g_h \neq g_f$ [h has a higher gain]

or (ii) $g_h = g_f$, $v_h \geq v_f$ and $v_h \neq v_f$ [h has the same gain but a higher bias]

or (iii) the policies f and h are equal.

Our proof if rather similar to the one in Derman [2] for the standard policy improvement step.

From the construction of the policy h we have the following two equations

$$P_{\tau,h} g_f = g_f + \psi \tag{5.1}$$

$$r_{\tau,h} + P_{\tau,h} v_f - Q_{\tau,h} P_{\tau,h} g_f = v_f + \gamma \ . \tag{5.2}$$

And further we have from lemma 4.3

$$P_{\tau,h} g_h = g_h \tag{5.3}$$

$$r_{\tau,h} + P_{\tau,h} v_h - Q_{\tau,h} P_{\tau,h} g_h = v_h \ . \tag{5.4}$$

Subtracting (5.1) from (5.3) and (5.2) from (5.4), writing $\Delta g = g_h - g_f$ and $\Delta v = v_h - v_f$ we obtain

$$P_{\tau,h} \Delta g = \Delta g - \psi \tag{5.5}$$

$$P_{\tau,h} \Delta v - Q_{\tau,h} P_{\tau,h} \Delta g = \Delta v - \gamma \ . \tag{5.6}$$

In order to prove $\Delta g \geq 0$ we need two additional lemmas.

Lemma 5.1.

(i) $\psi \geq 0$

(ii) if $\psi(i) = 0$ then $\gamma(i) \geq 0$.

Proof. (i) : Immediate from $g_f + \psi = P_{\tau,h} g_f \geq P_{\tau,f} g_f$ and $P_{\tau,f} g_f = g_f$

(ii): $\psi(i) = 0$ hence the rhs of (3.2) equals $g_f(i)$. From $\psi \geq 0$ and $P_f g_f = g_f$ the lhs of (3.2) with $a = f(i)$ is equal to

$$g_f(i) + \sum_{j \notin T_i} p(j|i,f(i))\psi(j) \geq g_f(i) .$$

But from (3.7) we see that the lhs of (3.2) is at most equal to the rhs. Therefore $f(i) \in A(i,f)$ and $\psi(j) = 0$ for all j with $p(j|i,f(i)) > 0$. Thus, if one starts in a state i with $\psi(i) = 0$ and plays f, then, with probability 1, one does not reach any state j with $\psi(j) > 0$ before τ. So, let $\hat{f}$ be any policy which prescribes $f(i)$ in the states with $\psi(i) = 0$. Then for all i with $\psi(i) = 0$

$$(v_f + \gamma)(i) \geq [r_{\tau,\hat{f}} + P_{\tau,\hat{f}} v_f - Q_{\tau,\hat{f}}(g_f + \psi)](i)$$

$$= [r_{\tau,f} + P_{\tau,f} v_f - Q_{\tau,f}(g_f + \psi)](i) = v_f(i) .$$

Hence $\gamma(i) \geq 0$. □

Two more notations. We will write $R_{\tau,h}$ for the set of states which are recurrent under $P_{\tau,h}$ and $P^*_{\tau,h}$ for the matrix

$$\lim_{N\to\infty} N^{-1} \sum_{n=0}^{N-1} P^n_{\tau,h} .$$

Lemma 5.2.

(i) On the set $R_{\tau,h}$ we have $\psi = 0$.

(ii) Δg is constant on each class of $P_{\tau,h}$.

Proof. (i): If we multiply (5.5) by $P^*_{\tau,h}$ and use $P^*_{\tau,h} P_{\tau,h} = P^*_{\tau,h}$ then we get $P^*_{\tau,h}\psi = 0$. So, with $\psi \geq 0$ we get $\psi = 0$ on $R_{\tau,h}$.

(ii): Substituting the result from (i) in (5.5) yields $P_{\tau,h}\Delta g = \Delta g$ on

$R_{\tau,h}$, hence also $P^n_{\tau,h}\Delta g = \Delta g$ and

$$\lim_{N\to\infty} N^{-1} \sum_{n=0}^{N-1} P^n_{\tau,h} g = P^*_{\tau,h}\Delta g = \Delta g.$$

So Δg must be constant on each class of $P_{\tau,h}$. □

Now we are ready to prove our first result

Lemma 5.3. On $R_{\tau,h}$ we have $\Delta g \geq 0$.

Proof. If we multiply (5.6) by $P^*_{\tau,h}$ we get

$$P^*_{\tau,h} Q_{\tau,h} P_{\tau,h} \Delta g = P^*_{\tau,h} \gamma \ . \tag{5.7}$$

From lemmas 5.2(i) and 5.1(ii) we have $\psi = 0$ hence $\gamma \geq 0$ on $R_{\tau,h}$. So $P^*_{\tau,h}\gamma \geq 0$. Moreover by lemma 5.2(ii) and corollary 4.2 we have $P_{\tau,h}\Delta g$ constant on each class of P_h. Now assume $\Delta g \equiv \sigma < 0$ on some class of $P_{\tau,h}$ then $P_{\tau,h}\Delta g \equiv \sigma$ on some class of P_h. So with $\tau \geq 1$ and $Q_{\tau,h}(i,j) = 0$ if j not in the same irreducible class of P_h we get $Q_{\tau,h}P_{\tau,h}\Delta g \leq \sigma < 0$ some class of $P_{\tau,h}$. Hence $P^*_{\tau,h}Q_{\tau,h}P_{\tau,h}\Delta g \leq \sigma < 0$ on some class of $P_{\tau,h}$. But $P^*_{\tau,h}\gamma \geq 0$ on S. Contradiction, so we must have $\Delta g \geq 0$ on $R_{\tau,h}$. □

In order to show $\Delta g \geq 0$ on S we use the following lemma. We write v_{min} for $\min_{i\in S} v(i)$.

Lemma 5.4. The set $D := \{i \in S \mid \Delta g(i) = \Delta g_{min}\}$ is closed under $P_{\tau,h}$.

Proof. From (5.5) we have for $i \in D$ the inequality

$$(P_{\tau,h}\Delta g)(i) = (\Delta g - \psi)(i) \leq (\Delta g)(i) = \Delta g_{min} \ .$$

But clearly $(P_{\tau,h}\Delta g)(j) \geq \Delta g_{min}$ for all $j \in S$. So we must have $(P_\tau \Delta g)(i) = \Delta g_{min}$, $i \in D$ and D is closed under $P_{\tau,h}$. □

From lemmas 5.3 and 5.4 we now have

Theorem 5.5. If h is a policy obtained from f by means of the modified policy iteration step then the gain of h is at least equal to the gain of f: $g_h \geq g_f$.

Proof. The set D of lemma 5.4 is closed under $P_{\tau,h}$ therefore contains an irreducible class of $P_{\tau,h}$ on which we have $\Delta g \equiv \Delta g_{min}$. But on $R_{\tau,h}$ we have $\Delta g \geq 0$ by lemma 5.3. Hence $\Delta g_{min} \geq 0$ or $g_h \geq g_f$ on S. □

Next we will show that $\Delta g = 0$ $[g_h = g_f]$ implies $v_h \geq v_f$. Substitution of $\Delta g = 0$ in (5.5) yields $\psi = 0$. Thus, by lemma 5.1(ii), $\gamma \geq 0$. And $\Delta g = 0$ reduces (5.6) to

$$P_{\tau,h}\Delta v = \Delta v - \gamma \ . \tag{5.8}$$

Multiplying (5.8) by $P^*_{\tau,h}$ gives $P^*_{\tau,h}\gamma = 0$, hence

Lemma 5.6. If $\Delta g = 0$ then $\gamma = 0$ on $R_{\tau,h}$.

Now we are able to prove the following important result. We write R_h for the set of recurrent states of P_h.

Lemma 5.7. If $\Delta g = 0$ then $h(i) = f(i)$ for all $i \in R_h$.

Proof. From the definition of h we see that it is sufficient to prove $f(i) \in B(i,f)$ for all $i \in R_h$. Let i be such that $\gamma(i) = 0$, then the rhs of (3.4) equals $v_f(i)$. From $\psi = 0$, $\gamma \geq 0$ and $r_f + P_f v_f - g_f = v_f$ the lhs of (3.4) with $a = f(i)$ becomes

$$v_f(i) + \sum_{j \notin T_i} p(j|i,f(i))\gamma(j) \geq v_f(i) \ .$$

But from (3.10) the lhs is at most equal to the rhs. So we have $f(i) \in B(i,f)$ and $\gamma(j) = 0$ for all j with $p(j|i,f(i)) > 0$.
From lemma 5.6 we have $\gamma = 0$ on $R_{\tau,h}$, hence $h(i) = f(i)$ for all $i \in R_{\tau,h}$. Further for all j for which there exists an i with $\gamma(i) = 0$ and $p(j|i,f(i)) > 0$ we have $\gamma(j) = 0$ hence $h(j) = f(j)$. Continuing to reason in this way we get $h(i) = f(i)$ for all i that can be reached from $R_{\tau,h}$ under h which is precisely the set R_h. □

Corollary 5.8. If $g_h = g_f$ then $v_h(i) = v_f(i)$ for all $i \in R_h$.

Proof. If we restrict $(1,1;h)$ to R_h then the solution is again unique and equal to the restriction of (g_h, v_h) to R_h. Since $f = h$ on R_h we must have $v_f = v_h$ on R_h. □

And finally we have

Theorem 5.9. If $g_h = g_f$ then $v_h \geq v_f + \gamma$ and if $\gamma = 0$ then $h = f$.

Proof. Analogous to the way we showed $\Delta g_{min} \geq 0$ (theorem 5.5) one proves $\Delta v_{min} = 0$, so $\Delta v \geq 0$, which substituted in (5.8) yields $\Delta v \geq \gamma$ or $v_h \geq v_f + \gamma$. Further if $\psi = \gamma = 0$ then it is immediately clear from (3.1) and (3.3) that $f(i) \in B(i,f)$ for all $i \in S$, hence $h = f$. □

6. The modified policy iteration algorithm yields a gain optimal policy

We still have to check whether the replacement in the policy iteration algorithm of the standard policy improvement step by the stopping time-based policy improvement step gives a convergent algorithm and produces an average optimal policy.

But this is not difficult. Clearly the modified policy iteration algorithm must converge as each improvement yields a new policy and there are only finitely many policies. Further let h^* be a policy to which the algorithm has converged then we already know from section 2 that h^* must be optimal gain.

A different way to see that h^* must be average optimal is obtained from section 5 as follows.

Let h be an arbitrary policy and let h^* replace f in all places in section 5. Then again we can write down the equations (5.1) - (5.6). But now $\psi \leq 0$ and if $\psi(i) = 0$ then $\gamma(i) \leq 0$ and lemma 5.2 holds as well. Anew we obtain (5.7) but now $P^*_{\tau,h}\gamma \leq 0$, so with lemma 5.3 we have $\Delta g \leq 0$ on $R_{\tau,h}$. With the analogon of lemma 5.4, $D' := \{i \in S \mid \Delta g(i) = \Delta g_{max}\}$ is closed under $P_{\tau,h}$, we get $\Delta g \leq 0$ on S.

. Remarks and extensions

(i) That our stopping times were transition memoryless turned out to be important in lemmas 3.1 and 3.2.
For other stopping times the restriction to policies is not allowed in general (cf. theorem 3.3 in Wessels [11]).
We also used $\tau \geq 1$ in lemma 5.3. The fact that τ is finite has been used throughout [for example $(I - \bar{P}_f)^{-1}$ and all strategies being equalizing].

(ii) For special transition memoryless stopping times [e.g. Gauss-Seidel] the policy improvement stop can be performed componentwise and then the amount of work will be the same as for the standard one. The improvement step can not be performed componentwise if τ is such that a path $(i_0,\ldots,i_n,\ldots)$ with $\tau(i_0,\ldots) > n$, $i_0 = i_n$ and $i_k \neq i_0$ for some $0 < k < n$ may occur; the case of a real cycle.

(iii) Here we considered nonrandomized stopping times only, the extension to randomized stopping times however is straightforward (cf. Van Nunen [8]).

(iv) We have made the restriction to finite stopping times. From a numerical point of view [compare (ii)] the only relevant case where $\mathbb{P}_{i,\pi}(\tau = \infty) > 0$ will occur is the Jacobi or Jacobi + Gauss-Seidel step [i.e. the cases $T = \{(i,j) | j \neq i\}$ or $T = \{(i,j) | j > i\}$] with $p(i|i,a) = 1$ for some i,a. Then $\mathbb{P}_{i,f}(\tau = \infty) = 1$ for all f with $f(i) = a$.
Our approach depended heavily on the finiteness of τ, but it seems possible to extend the results of this paper with some modifications to the case of nonfinite τ. A much simpler approach that would give us almost the Jacobi step would be to consider randomized stopping times where the probability of not stopping after (i,i) is though very close but not equal to 1.

(v) In section 2 we only considered the first two terms of the Laurent series expansion for $v_{\beta,f}$

$$v_{\beta,f} = \sum_{n=1}^{\infty} (1 - \beta)^n z_n(f) ,$$

with $z_{-1}(f) = g_f$ and $z_0(f) = v_f$ (cf. Miller and Veinott [7]). In a companion paper we will consider the whole series and obtain similar results as in Miller and Veinott. For example taking 3 instead of 2

terms of the Laurent series yields a stopping time-based improvement step which produces a bias optimal strategy.

References

[1] Blackwell. D., Discrete dynamic programming. Ann. Math. Statist. 33 (1962), 719-726.

[2] Derman, C., Finite state Markovian decision processes, Academic Press New York etc., (1970).

[3] Hastings, N.A.J., Some notes on dynamic programming and replacement, Operations Res. Quart. 19 (1968), 453-464.

[4] van Hee, K.M. and J. Wessels, Markov decision processes and strongly excessive functions, Stoch. Proc. and their Appl., to appear.

[5] Hordijk, A., Convergent dynamic programming, Stanford, Dept. Operations Res., Stanford University, 1974 (Techn. Rep. 28).

[6] Howard, R.A., Dynamic programming and Markov processes, Cambridge (Mass.), M.I.T. Press, 1960.

[7] Miller, B.L. and A.F. Veinott, Discrete dynamic programming with a small interest rate, Ann. Math. Statist. 40 (1969), 366-370.

[8] van Nunen, J.A.E.E., Contracting Markov decision processes, Amsterdam. Mathematisch Centrum, 1976 (MC tract no. 71).

[9] van Nunen, J.A.E.E. and J. Wessels, A principle for generating optimization procedures for discounted Markov decision processes, Amsterdam, North-Holland Publ. Comp., 1974. Colloquia Societatis Janos Bolyai, Vol. 12, 683-695.

[10] van Nunen, J.A.E.E. and J. Wessels, Markov decision processes with unbounded rewards, Markov decision theory, ed. by H.C. Tijms & J. Wessels, Amsterdam, Mathematisch Centrum 1977 (MC tract no. 93), 1-24.

[11] Wessels, J., Stopping times and Markov programming, Transactions of the seventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Prague, 1977, vol. A, 575-585.