Stopping times and Markov programming

TECHNOLOGICAL UNIVERSITY EINDHOVEN

Department of Mathematics

STATISTICS AND OPERATIONS RESEARCH GROUP

memorandum COSOR 74-09

Stopping times and Markov programming

by

J. Wessels

Summary: Using stopping times, a class of successive approximation methods for discounted Markov decision problems is constructed. This class contains many known procedures and a number of new ones.

Eindhoven, June 1974

Stopping times and Markov programming

J. Wessels

Eindhoven

0. Introduction.

In this paper we consider finite state Markov decision processes with finite decision spaces for each state. The optimality criterion will be total expected discounted reward over an infinite time horizon. For these problems a great variety of optimization procedures has been developed. We divide them in two classes:
policy improvement procedures and
policy improvement-value determination procedures.
For procedures of the first class the main part of each iteration step is a policy improvement procedure ([1], [2], [3], [9]). For procedures of the second class each iteration step contains a policy improvement part and a part in which the values for the new strategy are estimated or computed ([4], [5], [6], [7], [8], [3]). As a matter of fact it is possible to expand any procedure of the first class to a procedure of the second class. For different procedures this has been proved in [3]. A general approach will be presented in a forthcoming paper.
In this paper a unifying approach will be given for all known policy improvement procedures. At the same time a number of new policy improvement procedures is generated. It is proved that, in a way, our generating principle is exhaustive.

We use stopping times for the generation of policy improvement procedures. Actually the choice of a stopping time (or equivalently a go ahead set) will determine a procedure. We will present sufficient and necessary conditions for the stopping time which guarantee the convergence of the procedure (non-zero stopping times) and which guarantee that the procedure only requires the use of stationary Markov or memoryless decision rules (stationary second order Markov or transition memoryless stopping times).
The main tool in this paper consists of the theory of monotone contraction mappings, which has been used intensively in the past in this type of problems ([10], [11]).

1. Stopping times and contraction in stochastic processes.

We consider a stochastic process in a finite state space $S := \{1,\ldots,N\}$ and in discrete time ($n = 0,1,2,\ldots$). S^∞ is the set of allowed paths.

Definition 1.1.

a. The function τ on S^∞ with integer values between 0 and ∞ (bounds included) is called a *stopping time* if and only if

$$\tau^{\leftarrow}(n) = B \times S^\infty, \text{ with } B \subset S^{n+1} .$$

b. A nonempty subset A of $\bigcup_{k=0}^{\infty} S^k$ is called a *go ahead set*, if and only if

$$(\alpha,\beta) \in A \Rightarrow \alpha \in A \text{ for all } \alpha,\beta \in \bigcup_{k=0}^{\infty} S^k .$$

(S^0 only contains a null-tuple which concatenates to α with any α; our definitions imply that any go ahead set contains this null-tuple).

Notations. $\equiv A_n = \bigcup_{k=0}^{n} S^k$ $(0 \le n \le \infty)$.

$\equiv$ the i-th component of $\alpha \in S^n$ $(n \ge 1)$ is denoted by $[\alpha]_{i-1}$;

$\equiv$ if $\alpha \in S^n$, k_α is defined to be n;

$\equiv$ hence $\alpha \in S^n$ $(n \ge 1)$ may be written as $([\alpha]_0,[\alpha]_1,\ldots,[\alpha]_{k_\alpha-1})$;

$\equiv$ hence $k_\gamma = k_\alpha + k_\beta$ if $\gamma = (\alpha,\beta)$;

$\equiv A(i) = \{\alpha \in A \mid [\alpha]_0 = i \text{ if } k_\alpha \ge 1\}$.

There is a one-to-one correspondence between stopping times and go ahead sets:

$$A = \bigcup_{n=0}^{\infty} \{\alpha \in S^n \mid \forall_{\beta \in S^\infty} \tau(\alpha,\beta) \ge n\} .$$

The correspondence between stopping times and go ahead sets may be represented by:

$$\alpha \in A,\ (\alpha,\ell) \notin A,\ \ell \in S \Leftrightarrow \tau(\alpha,\ell,s_1,s_2,\ldots) = k_\alpha ,$$

$$(\ell) \notin A,\ \ell \in S \Leftrightarrow \tau(\ell,s_1,s_2,\ldots) = 0 .$$

We will apply the concepts of stopping time and go ahead set at will, always with this correspondence in mind.

Definition 1.2. A stopping time τ (or its go ahead set A) is said to be *nonzero* if and only if $\tau(\alpha) \geq 1$ for all $\alpha \in S^\infty$, or (equivalently) $S \subset A$.

The only nonzero stopping time which is an entrytime (memoryless) is $\tau \equiv \infty$ ($A = A_\infty$).

Examples of nonzero stopping times.

1.1. A_n $(1 \leq n \leq \infty)$;

1.2. A_H defined by : $A_H(i) = S^0 \cup \{(i)\} \cup \{(i,\alpha) \mid \alpha \in \bigcup_{j=1}^{i-1} A_H(j)\}$;

1.3. A_R defined by : $A_R(i) = \bigcup_{n=0}^{\infty} \{\alpha \in S^n \mid [\alpha]_j = i,\ j = 0,\ldots,n-1, \text{ if } n \geq 1\}$;

1.4. A_E with E a subset of S, defined by:

$$A_E = \bigcup_{n=2}^{\infty} \{\alpha \in S^n \mid [\alpha]_j \in E,\ j = 1,\ldots,n-1\} \cup S \cup S^0$$

$(E = S \Rightarrow A_E = A_\infty;\ E = \emptyset \Rightarrow A_E = A_1)$.

We now suppose that a reward structure has been given: at each time instant n a reward is earned. This reward $q(\alpha)$ depends on the history until that time: $\alpha \in S^{n+1}$. So the reward structure is a function q on A_∞. q is supposed to be bounded and (without loss of generality) to be zero on S^0. Rewards are discounted with discountfactor β $(0 < \beta < 1)$. We further denote the state of the stochastic process at time n by the random variable x_n and reward at time n by the random variable q_n. The probability of path $\alpha \in S^n$ is denoted by $P(\alpha)$. $P(\alpha \mid i)$ denotes the probability of α given $x_0 = i$. All such conditional probabilities are supposed to be defined properly. Defined on the process, the stopping time is a random variable.

Definition 1.3. A is a go ahead set, τ its corresponding stopping time. The operator L_A (or L_τ) on R^N is defined by:

$$(L_\tau v)(i) = E\Big(\sum_{k=0}^{\tau-1} \beta^k q_k + \beta^\tau v(x_\tau) \mid x_0 = i\Big)$$

(where E denotes expectation), or equivalently:

$$(L_A v)(i) = \sum_{\alpha \in A(i)} P(\alpha|i)\beta^{k_\alpha - 1} q(\alpha) + \sum_{\substack{\alpha \in A(i) \\ \ell \in S \\ (\alpha,\ell) \notin A(i)}} P(\alpha,\ell|i)\beta^{k_\alpha} v(\ell).$$

Theorem 1.1. (τ is a stopping time).

a. L_τ is a monotone mapping: $v \geq w \Rightarrow L_\tau v \geq L_\tau w$;

b. L_τ is strictly contracting with respect to supnorm in R^N if and only if τ is nonzero;

c. the contraction radius ρ_τ (or ρ_A) lies for nonzero τ between 0 and β (bounds included):

$$\rho_\tau = \max_{i \in S} E(\beta^\tau | x_0 = i);$$

d. $\rho_A \leq \rho_B$ if A and B go ahead sets with $A \supset B$, or $\rho_\tau \leq \rho_\sigma$ if $\tau \geq \sigma$.

Proof. The proof follows straightforward.

A natural question arises after the observation that strictly contracting mappings on $\mathbb{R}^N$ possess a unique fixed point: which point is mapped on itself by L_τ if τ is nonzero?

Lemma 1.1. If the stochastic process $\{x_n | n = 0,1,\ldots\}$, the nonzero stopping time τ, and the reward function q satisfy

$$E(\beta^\tau q_{\tau+\ell} | x_0=i, \tau<\infty, x_\tau=j) = E(\beta^\tau | x_0=i, \tau<\infty, x_\tau=j) E(q_\ell | x_0=j) \text{ for } \ell=1,2,..,i,j \in S$$

with $P(x_0=i, \tau<\infty, x_\tau=j) > 0$,

then L_τ possesses the unique fixed point $L_{A_\infty} 0$ (where 0 denotes the null-vector in $\mathbb{R}^N$).

Proof. Since L_τ possesses a unique fixed point if τ is nonzero, we only have to verify whether:

$$L_\tau L_{A_\infty} 0 = L_{A_\infty} 0, \text{ where } L_{A_\infty} 0(i) = E\left(\sum_{\ell=0}^{\infty} \beta^\ell q_\ell | x_0=i\right).$$

Using

$$L_\tau v(i) = E\left(\sum_{k=0}^{\tau-1} \beta^k q_k | x_0=i\right) + \sum_{j \in S} P(\tau<\infty, x_\tau=j | x_0=i) E(\beta^\tau | x_0=i, \tau<\infty, x_\tau=j) v(j)$$

the proof follows straightforward.

Theorem 1.2. Suppose:

1) the stopping time τ is nonzero,

2) the random variables τ and $q_{\tau+\ell}$ are conditionally independent (condition:

$$x_0 = i,\ \tau < \infty,\ x_\tau = j) \text{ for all } \ell \in N,\ i,j \in S,$$

3) $E(q_{\tau+\ell} \mid x_0 = i,\ \tau < \infty,\ x_\tau = j) = E(q_\ell \mid x_0 = j)$ for all $\ell \in N,\ i,j \in S,$

then $L_{A_\infty} 0$ is the unique fixed point of L_τ.

Proof. $E(\beta^\tau q_{\tau+\ell} \mid x_0 = i,\ \tau < \infty,\ x_\tau = j) =$

$$E(\beta^\tau \mid x_0 = i,\ \tau < \infty,\ x_\tau = j)E(q_{\tau+\ell} \mid x_0 = i,\ \tau < \infty,\ x_\tau = j) =$$

$$E(\beta^\tau \mid x_0 = i,\ \tau < \infty,\ x_\tau = j)E(q_\ell \mid x_0 = j)\ .$$

The statement is now implied by the foregoing lemma.

Corollary 1.2. If $\{x_n \mid n = 0,1,2,\ldots\}$ is a homogeneous Markov chain and $q(\alpha) = r([\alpha]_{k_\alpha - 1})$ for $k_\alpha \geq 1$, then $L_{A_\infty} 0$ is the unique fixed point for L_τ with τ a nonzero stopping time.

Stopping times and contraction in Markov decision processes.

From this section on we will treat Markov decision processes with state space S as described below.

Definition 2.1.

- a *decision rule* D is a function on $\bigcup_{k=1}^{\infty} S^k$ with values in a given set K (here supposed to be finite and nonempty);
- the decision rule D is said to be *memoryless* (stationary Markov) if $D(\alpha) = D([\alpha]_{k_\alpha - 1})$ for each $\alpha \in \bigcup_{k=1}^{\infty} S^k$;
- the set of decision rules is denoted by $\mathcal{D}$, the set of memoryless decision rules by M;
- the sequence $\{D_n\}_{n=1}^{\infty}$ of decision rules is said to converge to $D \in \mathcal{D}$, if and only if for each $\alpha \in \bigcup_{k=1}^{\infty} S^k$ holds: $\lim_{n\to\infty} D_n(\alpha) = D(\alpha)$.

Lemma 2.1. $\mathcal{D}$ is compact in the topology induced by the limit concept in $\mathcal{D}$.

Proof. The proof proceeds exactly as in [12]: the topology induced by the limit concepts of componentwise convergence is the same as the product topology in the countably infinite topological product of the sets of all maps of S^n into K (n = 1,2,...) (see e.g. Kelley [13]). Hence the compactness of $\mathcal{D}$ follows by Tychonov's theorem.

We suppose that each decision rule D determines a probability law for the stochastic process $\{x_n | n=0,1,...\}$ in the following way:

$$P_D(\alpha,j,\ell|i) = P_D(\alpha,j|i)p_{j\ell}^{D(\alpha,\ell)} \text{ for } \alpha \in A_\infty,\ i,j,\ell \in S,$$

here $p_{j\ell}^k$ $(j,\ell \in S,\ k \in K)$ are supposed to be given numbers satisfying:
$p_{j\ell}^k \geq 0$, $\sum_{\ell\in S} p_{j\ell}^k = 1$,

A visit to state i with a decision k gives the reward r_i^k.

Under these conditions a decision rule determines a stochastic process on S with the reward function q defined by:

$$q(\alpha,\ell) = r_{(\ell)}^{D(\alpha,\ell)} \qquad (\alpha \in A_\infty,\ \ell \in S).$$

Since we wish to consider such processes for different decisions rules D, we use D as a subscript or a superscript: $P_D(\alpha)$, L_τ^D, $E_D(\beta^\tau|x_0=i)$, etc.

Lemma 2.2. If D_1 and D_2 coincide on A: $L_A^{D_1} = L_A^{D_2}$.

Lemma 2.3. $L_\tau^D v$ is a continuous function of D (τ,v fixed).

From the foregoing section (theorem 1.1), it follows that L_τ^D is monotone and (if τ is nonzero) strictly contracting with contraction radius:

$$\rho_\tau^D = \max_{i\in S} E_D(\beta^\tau|x_0=i) .$$

The following theorem gives an assertion about the fixed point.

Theorem 2.1.

≡ If D is memoryless, then L^D has the fixed point $L_{A_\infty}^D 0$ for all nonzero τ;

≡ if D is not memoryless, then there exists a situation $(\{p_{ij}^k, r_{(i)}^k\})$ such that $L_{\tau_1}^D$ and $L_{\tau_2}^D$ possess different fixed points for certain τ_1 and τ_2.

Proof.

$\Rightarrow$ If D is memoryless, then $D(\alpha,\ell) = D(\alpha)$. Hence the process $\{x_n | n\geq 0\}$ then becomes a Markov chain with rewards only depending on the current state. Corollary 1.2 now produces the result.

$\Rightarrow$ D not memoryless implies: $\#\ K > 1$. τ_1 and τ_2 may be chosen identical to 1 and ∞ respectively.

Now: $L^D_{\tau_1}$ possesses the fixed point $L^{D_0}_{A_\infty}0$ with $D_0(\alpha,\ell) := D(\ell)$ for any $\alpha \in A_\infty$, $\ell \in S$;

$L^D_{\tau_2}$ possesses the fixed point $L^D_{A_\infty}0$.

It is not difficult to find p^k_{ij}'s and $r^k_{(i)}$'s, such that the two fixed points are different: $r^k_{(i)} = \delta_{kD(i)}$ for all $i \in S$, $k \in K$, then $(L^{D_0}_{A_\infty}0)(i) = (1-\beta)^{-1}$, while $(L^D_{A_\infty}0)(i) < (1-\beta)^{-1}$ for at least one $i \in S$, if $p^{D(j)}_{j\ell} > 0$ for all $j,\ell \in S$.

Example 2.1. For A_R we get:

$$\rho^D_{A_R} = \max_i[\beta(1-p^{D(i)}_{ii}) + \beta^2 p^{D(i)}_{ii}(1-p^{D(i,i)}_{ii}) + \ldots] \leq \beta\,\frac{1-p}{1-\beta p}, \text{ with } p := \min_{i,k} p^k_{ii}\ .$$

If D memoryless: $\rho^D_{A_R} = \beta\,\frac{1-q}{1-\beta q}$, with $q := \min_i p^{D(i)}_{ii}$.

Lemma 2.4. (τ is an arbitrary stopping time).
For any $v \in R^N$, there exists a decision rule D_0, such that $L^{D_0}_\tau v \geq L^D_\tau v$ (componentwise) for all D.

Notation. The vector $L^{D_0}_\tau v$ of the foregoing lemma, will be denoted by

$$\max_D L^D_\tau v\ ,\ U_\tau v,\ \max_D L^D_A v\ ,\ U_A v.$$

. A set of optimization procedures.

The operator U_τ serves for some specific choices of τ to construct optimization procedures, which aim actually at finding $U_{A_\infty}0$. In the set-up of this paper the question now arises how generally it is true that U_τ induces a procedure.

Theorem 3.1.

a. The operator U_τ on R^N is strictly contracting, if and only if τ is nonzero;

b. the contraction radius ν_τ of U_τ satisfies: $\nu_\tau = \max_D \rho_\tau^D$;

c. for all nonzero τ the operators U_τ posses the fixed point $U_{A_\infty} 0$;

d. $\nu_{\tau_1} \leq \nu_{\tau_2}$ if $\tau_1 \geq \tau_2$.

Proof.

a. Suppose $L_\tau^{D_1} v = U_\tau v$, $L_\tau^{D_2} w = U_\tau w$.

$$L_\tau^{D_2} v - L_\tau^{D_2} w \leq U_\tau v - U_\tau w \leq L_\tau^{D_1} v - L_\tau^{D_1} w .$$

This implies (as in theorem 1.1):

$$\| U_\tau v - U_\tau w \| \leq \| v - w \| \max\{\rho_\tau^{D_1}, \rho_\tau^{D_2}\} \leq \| v - w \| \max_D \rho_\tau^D .$$

The existence of the last maximum is a consequence of lemma 2.4 (with $r_i^k \equiv 0$).

Hence $\nu_\tau \leq \max_D \rho_\tau^D$, which already implies the strict contractness property of U_τ for nonzero τ. That U_τ, if τ is not nonzero, is not strictly contracting follows easily form the fact that any such τ possesses an $i \in S$ with $(U_\tau v)(i) = v(i)$.

b. Take $v(i) = V > 0$, $w(i) = 0$ (all $i \in S$), then:

$$(U_\tau v)(i) - (U_\tau w)(i) = \max_D \left[(L_\tau^D 0)(i) + V E_D(\beta^\tau | x_0 = i) \right] - \max_D (L_\tau^D 0)(i)$$

$$= V \{ \max_D \left[\frac{1}{V} (L_\tau^D 0)(i) + E_D(\beta^\tau | x_0 = i) \right] - \max_D \frac{1}{V} (L_\tau^D 0)(i) \}.$$

Choose $V > 0$, such that $\frac{1}{V} \| L_A^D 0 \| < \varepsilon$ for all D ($\varepsilon > 0$).

Then $\| U_\tau v - U_\tau w \| \geq V \{ -\varepsilon + \max_D \max_i E_D(\beta^\tau | x_0 = i) - \varepsilon \}$.

Hence $\nu_\tau \geq \max_D \rho_\tau^D$.

c. Suppose A is nonzero, hence U_A posses a unique fixed point.

Consider $U_A U_{A_\infty} 0$.

Its i-th component is equal to:

$$\max_D \Big[\sum_{\alpha\in A(i)} \beta^{k_\alpha - 1} P_D(\alpha|i) r^{D(\alpha)}_{([\alpha]_{k_\alpha - 1})} + \sum_{(\alpha,\ell)\in B} P_D(\alpha,\ell|i)\beta^{k_\alpha} \cdot$$

$$\{\max_{D_1} \sum_{\gamma\in A_\infty(\ell)} \beta^{k_\gamma - 1} P_{D_1}(\gamma|\ell) r^{D_1(\gamma)}_{([\gamma]_{k_\gamma - 1})}\}\Big] =$$

$$= \max_{D,D_1} \Big[\sum_{\alpha\in A(i)} \beta^{k_\alpha - 1} P_D(\alpha|i) r^{D(\alpha)}_{([\alpha]_{k_\alpha - 1})} + \sum_{(\alpha,\ell)\in B} \beta^{k_\alpha + k_\gamma - 1} \cdot$$

$$P_D(\alpha,[\gamma]_0|i) P_{D_1}(\gamma|[\gamma]_0) r^{D_1(\gamma)}_{([\gamma]_{k_\gamma - 1})}\Big] =$$

$$= \max_{D,D_1} (L^{(D,D_1)}_{A_\infty} 0)(i) = \max_D (L^D_{A_\infty} 0)(i) ,$$

where (D,D_1) denotes the decision rule defined by: $(D,D_1)(\alpha) = D(\alpha)$, if $\alpha \in A$, $(D,D_1)(\alpha,\gamma) = D_1(\gamma)$, if $\alpha \in A$, $(\alpha,[\gamma]_0) \notin A$ and B contains the elements (α,γ) with $\alpha \in A(i)$, $\gamma \in A_\infty$, $(\alpha,[\gamma]_0) \notin A$, $\ell \in S \subset A_\infty$.
The last equality holds since the class of decision rules $\{(D,D_1)\}$ contains M and

$$\max_D L^D_{A_\infty} 0 = \max_{D\in M} L^D_{A_\infty} 0 \text{ (e.g. [6], [10], [12])}.$$

Examples 3.1. $\nu_{A_k} = \beta^k$ $(1 \le k \le \infty)$;

3.2. $\nu_{A_H} = \beta$;

3.3. $\nu_{A_R} = \beta \frac{1-p}{1-\beta p}$.

In principle this theorem makes it possible to construct an infinite number of procedures for finding $U_{A_\infty} 0$, namely choose a nonzero stopping time τ, choose a starting guess $v_0 \in R^N$, and define:

$$v_n = U_\tau v_{n-1} \quad (n = 1,2,\ldots) .$$

Then v_n converges to $U_{A_\infty} 0$.
Regrettably the computation of $U_\tau v$ may be equally cumbersome as the computation of $U_{A_\infty} 0$. So the following problem for investigation is the character-

ization of the nonzero stopping times which allow easy computation of $U_\tau v$. The two theorems in the sequel of this section provide the main step for such a characterization.

Definition 3.1. A stopping time τ (and its corresponding go ahead set A) is said to be *transition memoryless*, if and only if there is a subset T of S^2 and a subset S_0 of S, such that:

$$\tau(\alpha) = 0 \Leftrightarrow [\alpha]_0 \in S_0 ,$$

$$\tau(\alpha) = n > 0 \Leftrightarrow ([\alpha]_k,[\alpha]_{k+1}) \notin T \text{ for } k=0,..,n-2,\ ([\alpha]_{n-1},[\alpha]_n) \in T.$$

Lemma 3.1. Memoryless stopping times are transition memoryless.

Theorem 3.2. If τ is a transition memoryless stopping time:

$$U_\tau = \max_{D\in M} L_\tau^D .$$

Proof.
$$(L_\tau^D v)(i) = \begin{cases} v(i) \text{ if } i \in S_0 , \\ r_{(i)}^{D(i)} + \beta \sum_{j\in T(i)} p_{ij}^{D(i)} v(j) + \beta \sum_{\substack{j\notin T(i)\\ j\notin S_0}} p_{ij}^{D(i)} (L_\tau^{D_{ij}} v)(j) + \\ \qquad + \beta \sum_{\substack{j\notin T(i)\\ j\in S_0}} p_{ij}^{D(i)} (L_{\tau_1}^{D_{ij}} v)(j) \text{ if } i \notin S_0. \end{cases}$$

Here: $T(i) = \{j \in S \mid (i,j) \in T\}$,
D_{ij} is a decision rule with $D_{ij}(\alpha) = D(i,\alpha)$ if $[\alpha]_0 = j$,
τ_1 is the transition memoryless stopping time with the same T as τ, but with an empty S_0.

It is possible to define a new Markov decision process with essentially the same decision rules in such a way that $L_\tau^D v$ is exactly the vector of total expected discounted rewards. Hence for this new Markov decision process attention may be restricted to memoryless strategies (e.g. [6], [10], [12]), which implies the same for the original problem. This new Markov decision process is defined in the following way: $\bar{S}$, the new set of states, consists of S_0 and two representations of S: $S^* = \{s^* \mid s \in S\}$ and $S_* = \{s_* \mid s \in S\}$. So some states of S are three times represented in $\bar{S}$ and others two times.

For the states $\bar{s} \in S_* \cup S_0$ we define: $p^k_{\bar{s}\bar{s}} = 1$, $r^k_{(\bar{s})} = (1-\beta)v(s)$ $\quad (k \in K)$.

For the states $s^* \in S^*$ we define: $p^k_{s^*s_1^*} = p^k_{ss_1}$ if $(s,s_1) \notin T$ $\quad (k \in K)$,

$p^k_{s^*s_{1*}} = p^k_{ss_1}$ if $(s,s_1) \in T$ $\quad (k \in K)$,

$r^k_{(s^*)} = r^k_{(s)}$ $\quad (k \in K)$.

Transition memoryless stopping times are the only stopping times, for which restriction to memoryless decision rules is always allowed:

Theorem 3.3. Suppose the stopping time τ for the state set S is not transition memoryless, then there exists a Markov decision process with state set S (i.e. there exists a set K, and numbers $\{p^k_{ij}\},\{r^k_{(i)}\}$) such that for this Markov decision process

$$U_\tau \neq \max_{D \in M} L^D_\tau .$$

Remark. In fact, $\max_{D \in M} L^D_\tau$ may not be defined.

Proof. Representing τ by its go ahead set A, its not being transition memoryless implies the existence of a state i and two paths α,γ such that $B = \{j \mid (\alpha,i,j) \in A\} \neq \{j \mid (\gamma,i,j) \in A\} = C$, while (α,i), $(\gamma,i) \in A$.
This implies that in determining U_A the following forms have to be maximized with respect to k and ℓ respectively:

$$(*) \qquad r^k_{(i)} + \beta \sum_{j \notin B} p^k_{ij} v(j) + \beta \sum_{j \in B} p^k_{ij} (U_{\tau_1} v)(j) ,$$

$$(**) \qquad r^\ell_{(i)} + \beta \sum_{j \notin C} p^\ell_{ij} v(j) + \beta \sum_{j \in C} p^\ell_{ij} (U_{\tau_2} v)(j) ,$$

where $\tau_1(\delta) = \tau(\alpha,i,\delta)$ and $\tau_2(\delta) = \tau(\gamma,i,\delta)$.
By investigating different possibilities for the relation between B and C examples can be constructed for which the maximizing k and ℓ in (*) and (**) are different.

Conclusion. It will be clear that the existing policy improvement procedures [1], [8] follows directly from our unifying approach by choosing the corresponding stopping time. While the policy improvement procedure introduced by Reets [2] can be achieved by a slight extension of the set of allowed stopping times.

References

[1] J. MacQueen, A modified dynamic programming method for Markovian decision problems, J. Math. Anal. Appl. 14 (1966) 38-43.

[2] D. Reetz, Solution of a Markovian decision problem by successive over-relaxation, Z.f. Oper.Res. 17 (1973) 29-32.

[3] J.A.E.E.v.Nunen, Improved successive approximation methods for discounted Markov decision processes, Memorandum COSOR 74-06, Technological University Eindhoven, April 1974 (Dept. of Math.).

[4] R.A. Howard, Dynamic programming and Markov processes, MIT-Press, Cambridge, 1960.

[5] G.T.de Ghellinck, G.D. Eppen, Linear programming solutions for separable Markovian decision problems, Man. Sc. 13 (1967) 371-394.

[6] J. Wessels, J.A.E.E.v.Nunen, Discounted semi- Markov decision processes: linear programming and policy iteration, to appear in Statistica Neerlandica.

[7] J.A.E.E.v.Nunen, A set of successive approximation methods for discounted Markovian decision problems, Memorandum COSOR 73-09, Technological University Eindhoven, September 1973 (Dept. of Math.).

[8] N. Hastings, Some notes on dynamic programming and replacement, Oper. Res. Q. 19 (1968) 453-464.

[9] H. Schellhaas, Zur Extrapolation in Markoffschen Entscheidingsmodellen mit Diskontierung, Preprint nr 84 (1973), Technische Hochschule Darmstadt.

[10] D. Blackwell, Discounted dynamic programming, Ann. Math. Statist. 36 (1965) 226-234.

[11] E.V. Denardo, Contraction mappings in the theory underlying dynamic programming, SIAM-Review 9 (1967) 165-177.

[12] J. Wessels, Decision rules in Markovian decision processes with incompletely known transition probabilities, Technological University Eindhoven, 1968.

[13] J.L. Kelley, General topology, New York, 1955.

[14] H. Mine, S. Osaki, Markovian decision processes, New York 1970.

Technological University Eindhoven
Department of Mathematics.