The generation of successive approximation methods for Markov decision processes by using stopping times

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 76-22

The generation of successive approximation methods for Markov decision processes by using stopping times

by

J.A.E.E. van Nunen and J. Wessels

Eindhoven, November 1976

The Netherlands

The generation of successive approximation methods for Markov decision processes by using stopping times

by

J.A.E.E. van Nunen and J. Wessels

Summary

In this paper we will consider several variants of the standard successive approximation technique for Markov decision processes. It will be shown how these variants can be generated by stopping times.
Furthermore it will be demonstrated how this class of techniques can be extended to a class of value oriented techniques. This latter class contains as extreme elements several variants of Howard's policy iteration method.
For all methods presented extrapolations are given in the form of MacQueen's upper and lower bounds.

1. Introduction

In [1] we introduced the standard successive approximation method for Markov decision processes with respect to the total expected reward criterion. In fact there exist some variants of this method. These variants differ in the policy improvement procedure: the standard procedure may be replaced by a Gauss-Seidel procedure (see e.g. Hastings [10], Kushner and Kleinman [13]), an overrelaxation procedure (see Reetz[7] and Schellhaas [9]) or some other variants (see Van Nunen [4]). In [3] it has been shown that such variants can be generated by stopping times. This approach has been generalized in [2]. In section 2 we will introduce the main idea of this approach.
Policy iteration -with its several variants- as introduced by Howard [12] is usually not viewed upon as a successive approximation technique. However, in [5] it has been shown to be an extreme element of a class of extended successive approximation techniques, the so-called value-oriented methods. This approach has been combined in [6] with the stopping time approach. In [2] a further generalization has been given (mainly with respect to the conditions). Value-oriented methods will be treated in section 3 .
Section 4 will be devoted to upper and lowerbounds for the techniques

presented in the earlier section. Furthermore some remarks on numerical aspects will be made.

In this paper we will use the same notations as in [1], however, in order to keep the proofs simple, we will work under somewhat stronger assumptions. In fact, our assumptions are the same as those in [2]. For details we will refer repeatedly to [2].

Assumptions. Our assumptions are the same as the assumptions in [1], with assumption 2.3 (i) replaced by

(a) $$\exists_{M>0} \forall_{f\in F} \| r(f) - \bar{r} \| \leq M ,$$

(b) $$\sup_{\pi\in M} \mathbb{E}_i^{\pi} \sum_{n=0}^{\infty} |\bar{r}(X_n)| < \infty \qquad \text{for all } i \in S .$$

These stronger assumptions make the spaces V^- and W^- superfluous.
As remarked in [1] (remark 5.1) one may replace $\bar{r}$ in the assumptions (and definition of V) by a vector b with $b - \bar{r} \in W$. We will do so in this paper in order to facilitate referring to [2].

2. Stopping times and successive approximations

In this section we will show that each stopping time characterized by a goahead function δ for the sequence $\{X_n\}_{n=0}^{\infty}$ induces an operator U_δ on V, such that U_δ is monotone and (usually) contracting.
Furthermore all these contracting operators on V have the same unique fixed point v^*. So we have for any $v_0 \in V$ and any δ:

$$v_n := U_\delta v_{n-1} \in V \quad \text{for } n = 1,2,\ldots$$

and $v_n \to v$.

Definition 2.1. A (randomized) *go ahead function* δ is a function which maps

$$G_\infty := \bigcup_{k=1}^{\infty} S^k \quad \text{into } [0,1].$$

By Δ we denote the set of all go ahead functions.

$1 - \delta(s_0,s_1,\ldots,s_n)$ will be interpreted as the probability to stop the process at time n, given that $X_0 = s_0,\ldots,X_n = s_n$ and the process has not been stopped earlier.

Definition 2.2.

(a) $\delta \in \Delta$ is said to be *nonrandomized* if $\delta(\alpha) \in \{0,1\}$ for all $\alpha \in G_\infty$;

(b) $\delta \in \Delta$ is said to be *nonzero* if $\delta(i) > \varepsilon > 0$ for some ε and all $i \in S$;

(c) $\delta \in \Delta$ is said to be *transition memoryless* if $\delta(\alpha)$ only depends on the last two entries of α, for those α with at least two entries and satisfying $\delta(s_0,\ldots,s_k) \neq 0$ for all $k < n$, if $\alpha = s_0,\ldots,s_n$.

So for a transition memoryless go ahead function the stopping probability only depends on the most recent transition. The relevance of this notion will become clear in the course of this section.

Examples 2.1. Below some examples of nonzero go ahead functions will be given. These examples will be used repeatedly in this paper.

(a) Define the go ahead function δ_n $(n = 1,2,\ldots)$ by $\delta_n(\alpha) := 1$ if α contains less than $n + 1$ entries, otherwise $\delta_n(\alpha) := 0$. The go ahead functions δ_n are nonrandomized, δ_n is only transition memoryless if $n = 1$.

(b) define δ_R by $\delta_R(s,s,\ldots,s) := 1$ for all s and all sequences of finite length, $\delta_R(\alpha) := 0$ otherwise.
δ_R is nonrandomized and transition memoryless.

(c) δ_H is defined by $\delta_H(s_0,\ldots,s_n) := 1$ if $s_0 < s_1 < \ldots < s_n$ (any n), otherwise $\delta_H(\alpha) := 0$.
δ_H is nonrandomized and transition memoryless.

(d) $\delta_r(i) = \frac{1}{2}$ for all $i \in S$, $\delta_r(\alpha) := 0$ elsewhere.
δ_r is transition memoryless.

Since we introduced a probabilistic go ahead concept, we have to incorporate it in the probability space and measure. Therefore we extend the space $(S \times A)^\infty$ (see [1] section 2) to $(S \times E \times A)^\infty$, with $E := \{0,1\}$. Furthermore the stochastic process $\{X_t,Z_t\}_{t=0}^\infty$ to $\{X_t,Y_t,Z_t\}_{t=0}^\infty$, where $Y_t = 0$ as long as the process may go ahead.
Now any starting state i, any go ahead function δ, and any decision rule π determine a probability measure on $(S \times E \times A)^\infty$ with the required properties in an obvious way (see [2] for details). This probability measure will be denoted by $\mathbb{P}_i^{\pi,\delta}$. Expectations will be denoted by $\mathbb{E}_i^{\pi,\delta}$. Note that $\mathbb{P}_i^{\pi,\delta}$ and $\mathbb{P}_i^{\pi}$ are equal for events which do not depend on the variables Y_t.
In fact the go ahead concept induces a stopping time

Definition 2.3. The random variable τ taking values in $\{0,1,\ldots,\infty\}$ is defined by

$$\tau = n \;:\Leftrightarrow\; Y_0 = \ldots = Y_{n-1} = 0 \text{ and } Y_n = 1$$

$$\tau = \infty \;:\Leftrightarrow\; Y_t = 0 \text{ for all } t = 0,1,\ldots .$$

τ is a randomized stopping time with respect to $X_0,X_1,\ldots$.
Now we will introduce our operators.

Definition 2.4. For each $\delta \in \Delta$ and each strategy (= nonrandomized decision rule) π the operator L_δ^π on V is defined by

$$L_\delta^\pi v := \mathbb{E}^{\pi,\delta}\Big[\sum_{k=0}^{\tau-1} r(X_k,Z_k) + v(X_\tau)\Big] \quad \text{for } v \in V, \text{ with } v(X_\tau) := 0$$

if $\tau = \infty$.

Lemma 2.1. L_δ^π is monotone and (for nonzero δ) strictly contracting on V. Therefore L_δ^π possesses a unique fixed point v_δ^π in V.

Proof. The contraction factor of L_δ^π is $\|P_\delta(\pi)\| =: \rho_\delta^\pi$, where $P_\delta(\pi)$ is the matrix with (i,j) entry $\mathbb{P}_i^{\pi,\delta}(\tau < \infty, X_\tau = j)$. $\rho_\delta^\pi < 1$ if and only if δ is nonzero.

Examples. Take for π an arbitrary stationary strategy (f,f,...).

(a) $$(L_{\delta\,1}^{\pi} v)(i) = r(i,f(i)) + \sum_j p^{f(i)}(i,j)v(j) \ ;$$

(b) $$(L_{\delta\,R}^{\pi} v)(i) = [1 - p^{f(i)}(i,i)]^{-1}[r(i,f(i)) + \sum_{j\neq i} p^{f(i)}(i,j)v(j)] \ ;$$

(c) $$(L_{\delta\,H}^{\pi} v)(i) = r(i,f(i)) + \sum_{jj<i} p^{f(i)}(i,j)(L_{\delta\,H}^{\pi} v)(j) + \sum_{j\geq i} p^{f(i)}(i,j)v(j) \ .$$

(d) $$(L_{\delta\,r}^{\pi} v)(i) = \tfrac{1}{2}v(i) + \tfrac{1}{2}[r(i,f(i)) + \sum_j p^{f(i)}(i,j)v(j)] \ .$$

(e) let δ be nonzero, then $v_\delta^\pi = v(\pi)$, independent of δ.

Remark 2.1. If π is a nonstationary strategy then there exist values for $\{p^a(i,j), r(i,a)\}$ and go ahead functions δ' and δ'' such that $v_{\delta'}^\pi \neq v_{\delta''}^\pi$ (see lemma 5.1.7 in [2]).

We now come to the operators U_δ

Definition 2.5. The operator U_δ on V is defined by

$$U_\delta v := \sup_\pi L_\delta^\pi v \ ,$$

where the supremum is taken componentwise.

Note that L has only been defined for strategies π, so the supremum is only taken over the strategies (= nonrandomized decision rules). Extension to the randomized decision rules would not affect the value of $U_\delta v$.

Theorem 2.1. Let $\delta \in \Delta$, then U_δ is monotone and (only for nonzero δ) strictly contracting with contraction radius $\nu_\delta := \sup_\pi \rho_\delta^\pi$. Therefore U_δ possesses

(for nonzero δ) a unique fixed point. v^* is fixed point for all U_δ with δ nonzero.

Proof. For details we refer to the proof of theorem 5.2.1 in [2]. With respect to the last statement we remark:

$$U_\delta v^* \geq L_\delta^\pi v^* \geq L_\delta^\pi v(\pi) = v(\pi) \quad \text{if } \pi = (f,f,\ldots) .$$

Since f may be chosen such that $v(\pi) \geq v^* - \varepsilon\mu$ ([1] theorem 3.1 (ii)), we obtain $U_\delta v^* \geq v^*$. If we had $U_\delta v^* > v^*$, then it would be possible to construct a strategy π' with $v(\pi') > v^*$.

This theorem serves as the basis for a δ-based successive approximation algorithm, since $v_n := U_\delta v_{n-1}$ converges in norm to v^* if $v_0 \in V$.
In the definition of U_δ we take the supremum over all strategies. One would naturally prefer to restrict oneself to Markov strategies and even use the algorithm for constructing ε-optimal stationary strategies. The following theorem (for the proof we refer to [2] theorem 5.2.2 and 5.2.3) shows that the concept of transition memoryless go ahead functions plays a crucial role in this problem.

Theorem 2.2.

(a) Let δ be transition memoryless, $\varepsilon > 0$, $v \in V$.
Then there exists a policy f, such that

$$L_\delta^f v \geq U_\delta v - \varepsilon\mu .$$

(b) Let δ be not transition memoryless, then there exist values for the parameters $\{p^a(i,j), r(i,a)\}$, such that for some $v \in V$ and some ε there is no $f \in F$ with

$$L_\delta^f v \geq U_\delta v - \varepsilon\mu .$$

Hence, if δ is transition memoryless we have

$$U_\delta v = \sup_f L_\delta^f v ,$$

where the sup is not necessarily componentwise. Whereas if δ is not transition memoryless

$$\sup_{f} L_{\delta}^{f} v$$

may not be defined.

For nonzero and transition memoryless go ahead functions we now obtain the following iteration procedures

(a) (if $\sup_{f} L_{\delta}^{f} v$ is attained for some f).

Choose $v_0 \in V$, define $v_n := U_\delta v_{n-1}$ and choose f_n such that $v_n = L_\delta^{f_n} v_{n-1}$, then

(i) $\|v_n - v^*\| \leq \nu_\delta^n \|v_0 - v^*\|$

(ii) $\|v_n - v(f_n)\| \leq (1 - \nu_\delta)^{-1} \nu_\delta \|v_n - v_{n-1}\|$

(iii) if v_0 satisfies $U_\delta v_0 \geq v_0$, then $v_{n-1} \leq v_n \leq v(f_n) \leq v^*$.

(b) Choose $\varepsilon > 0$ and $v_0 \in V$ with $v_0 \leq U_\delta v_0 - \varepsilon\mu$.

Choose f_n $(n = 1,\ldots)$ such that

$$L_\delta^{f_n} v_{n-1} \geq \max\{v_{n-1},\ U_\delta v_{n-1} - \varepsilon(1 - \nu_\delta)\mu\}$$

define

$$v_n := L_\delta^{f_n} v_{n-1} ,$$

then

(i) $\|v_n - v^*\| < \varepsilon$ for n sufficiently large

(ii) $v_{n-1} \leq v_n \leq v(f_n) \leq v^*$.

In fact, as in the case of δ_1, more efficient lower and upperbounds can be obtained (see section 4).

Examples 2.3. The examples 2.2 (a)-(b) induce numerically well-executable policy improvement procedures. In fact δ_1 induces the standard successive approximations technique based on Gauss-Jordan-iteration; δ_R induces Jacobi iteration (compare Porteus [14]); δ_H yields Gauss-Seidel iteration; other choices of δ yield overrelaxation and combinations of overrelaxation and Gauss-Seidel iteration (in this respect lemma 7.2.3 in [2] has interesting consequences).

3. Value oriented methods

In the foregoing section we developed a whole class of policy improvement procedures or successive approximations techniques. As we saw in section 2, at the n-th stage of any policy improvement procedure the best estimate for the optimal strategy is the stationary strategy f_n. This makes the next policy improvement more efficient if the value v_n is nearer to $v(f_n)$. In fact policy iteration techniques owe their high efficiency in the policy improvement part to the fact that they have $v_n = v(f_n)$. A disadvantage of policy iteration is in fact the computation of these v_n. However, there is an arbitrary way combining the advantages of policy iteration and successive approximations. Namely suppose that f_n is chosen such that

$$L_\delta^{f_n} v_{n-1} = U_\delta v_{n-1} ,$$

then define

$$v_n := (L_\delta^{f_n})^\lambda v_{n-1} \qquad (\lambda \in \{1,2,\ldots,\infty\}) .$$

Note that

$$\lim_{\lambda\to\infty} (L_\delta^{f_n})^\lambda v_{n-1} = v(f_n) ,$$

so by the choice of λ we in fact determine how good v_n approximates $v(f_n)$. The choice $\lambda = 1$ gives the successive approximation of section 2, whereas the choice $\lambda = \infty$ gives for any transition memoryless and nonzero go ahead function a variant of the policy iteration technique.
Below we give a more formal treatment.

Definition 3.1. Let δ be nonzero and transition memoryless and suppose that the $\sup_f L_\delta^f v$ is attained for some policy if $v \in V$. Furthermore we assume that we have a unique way of designating such a policy. We define the operators $U_\delta^{(\lambda)}$ on V for $\lambda = 1,2,\ldots,\infty$ by

$$U_\delta^{(\lambda)} v = (L_\delta^f)^\lambda v \ ,$$

if the sup in $U_\delta v$ is attained for f.
Note that

$$U_\delta^{(\infty)} v = \lim_{n\to\infty} (L_\delta^f)^n v = v(f) \ .$$

It does not seem revolutionary to conjecture that $v_n := U_\delta^{(\lambda)} v_{n-1}$ converges to v^* if $v_0 \in V$. However, one becomes somewhat more prudent as soon as one realizes that $U_\delta^{(\lambda)}$ is neither necessarily monotone, nor necessarily contracting as one can see in the following simple example for $\delta = \delta_1$, $S = \{1,2\}$, $\mu \equiv 1$, $A = \{1,2\}$: $p^1(i,2) = p^2(i,1) = 0.99$, $r(i,1) = 1$, other probabilities and rewards being zero.
Now one obtains for $v := (0,0)^T$, $w := (10,1)^T$ $\lim_{\lambda\to\infty} U_\delta^{(\lambda)} v = (100,100)^T$, whereas $\lim_{\lambda\to\infty} U_\delta^{(\lambda)} w = (0,0)^T$.

We will now prove that the proposed iteration step leads to a converging algorithm.

Theorem 3.1. Let the situation be such that $U_\delta^{(\lambda)}$ is defined and choose $v_0 \in V$ with $U_\delta v_0 \geq v_0$.
Then $v_n := U_\delta^{(\lambda)} v_{n-1}$ converges in norm to v^* ,
and

$$\| v_n - v^* \| \leq \nu_\delta^n \| v_0 - v^* \|$$

$$v_{n-1} \leq v_n \leq v(f_n) \leq v^* \ ,$$

where f_n is the policy (unique, possibly after tie breaking) which maximizes $L_\delta^f v_{n-1}$.

Proof. By assumption we have

$$L_\delta(f_1)v_0 = U_\delta v_0 \geq v_0 .$$

Hence

$$v_0 \leq L_\delta(f_1)v_0 \leq \ldots \leq [L_\delta(f_1)]\, v_0 = v_1 \leq \lim_{n\to\infty} [L_\delta(f_1)]^n v_0 = v(f_1) .$$

Since $U_\delta v_1 = L_\delta(f_2)v_1 \geq L_\delta(f_1)v_1$, one obtains $v_1 \leq v_2 \leq v(f_2)$.
By induction this gives $v_{n-1} \leq v_n \leq v(f_n) \leq v^*$. On the other hand $v_n \geq U_\delta^n v_0$, which tends to v^* for $n \to \infty$. Therefore $v_n \to v^*$ and

$$\|v_n - v^*\| \leq \|U_\delta^n v_0 - U_\delta^n v^*\| \leq \nu_\delta^n \|v_0 - v^*\| .$$

In the same way as in [1] for the standard algorithm one may obtain more sophisticated bounds (see section 4). Furthermore the assumption that the sup in $U_\delta v$ is attained can be weakened as in [1] by introducing approximations (in norm) of the sup. This can be extended in several ways. For a detailed description of these possibilities see [2].

As already stated, the case $\lambda = \infty$ represents a variety of policy iteration procedures. In fact the procedures (for any nonzero transition memoryless δ) generate sequences of policies with increasing value. Hence an optimal policy is obtained after a finite number of iterations if the state and action spaces are finite.
If $\delta = \delta_1$, then we have the standard policy iteration algorithm as introduced by Howard in [12] for the finite state, finite action discounted case. If $\delta = \delta_H$, then we have the Gauss-Seidel variant as introduced by Hastings [10].

4. Some remarks on numerical and other aspects

For the algorithms based on the operators U_δ (section 2) and $U_\delta^{(\lambda)}$ (section 3) we proved geometric convergence. However, the extrapolation based on the convergence rate only are usually not very good. As in the case of U_{δ_1} (see [1]) one can obtain better bounds rather easily. For the case the sup in $U_\delta v_n$ is attained and exactly computed in the algorithm based on $U_\delta^{(\lambda)}$ $(\lambda = 1,2,\ldots,\infty)$ we obtain, if $U_\delta v_0 \geq v_0$:

$$v_n + (1 - \rho_{f_{n+1}})^{-1} \| L_\delta(f_{n+1}) v_n - v_n \|_- \leq v(f_{n+1}) \leq v^* \leq$$

$$v_n + (1 - \nu_\delta)^{-1} \| L_\delta(f_{n+1}) v_n - v_n \| ,$$

where

$$\rho_f := \inf_f \mu^{-1}(i) \sum_j p^{f(i)}(i,j)\mu(j), \quad \| v \|_- := \inf_f \mu^{-1}(i) v(i) .$$

For a more detailed description we refer to [2]. The proof in this case is completely similar to the proof in the case $\delta = \delta_1$.

For numerical experience it appears that value oriented methods can give a considerable gain in computational efficiency. This is especially true if the policy improvement procedure requires many operations. Generally speaking one may say that U_δ-based successive approximations methods only need a small number of iterations to reach a near-optimal policy, however, the proof of this near-optimality requires relatively many additional iterations. So in quite a lot of iterations f_n does not change substantially. Therefore it is efficient to choose λ greater than one. In fact it is still more profitable to increase the value of λ in subsequent situations. To give an idea of the gain in computational efficiency we mention that we found in a number of examples with $\delta = \delta_1$ a saving in computing time of 20 - 40% when we took $\lambda = 5$ instead of $\lambda = 1$ (in both situations we used a suboptimality test; the numbers of states ranged between 40 and 1000), see [4].

In all procedures (all δ and all λ) the standard suboptimality test is allowed and also the more sophisticated and more efficient suboptimality test which is described in the paper by Hastings and Van Nunen [10] in this volume.

Instead of defining δ-based operators U_δ one may transform the data in the problem and solving the transformed problem by the standard successive approximation methods. This approach has been presented by Porteus in [14]. In our notation the transformation is

$$\tilde{r}(f) := \mathbb{E}^{f,\delta} \sum_{n=0}^{\tau-1} r(X_n, Z_n) ,$$

$$\tilde{P}(f) := P_{\delta}(f) \quad \text{(see proof of lemma 2.1)} .$$

By introducing the matrices $Q(f)$ with $Q(f)(i,j) := p^{f}(i,j)\delta(i,j)$ we obtain

$$\tilde{P}(f) = \sum_{k=0}^{\infty} Q^{k}(f)[P(f) - Q(f)] ,$$

$$\tilde{r}(f) = \sum_{k=0}^{\infty} Q^{k}(f)r(f) ,$$

being exactly Porteus' preinverse transformation. In fact we showed in section 2, that the transformed problem possesses the same optimal value vector as the original problem.

In fact some extension is possible with respect to the conditions unde which the U_{δ}- and $U_{\delta}^{(\lambda)}$-based procedures converge. We mentioned already the kind of conditions of [1]. Another approach is in considering a fixed δ and require strict or N-stage contraction for U_{δ} on V or W. In [8] Reetz chooses such an approach for $\delta = \delta_{H}$. One might conjecture that -as in the case of δ_{1} (see [1]) - N-stage contraction implies 1-stage contraction with respect to a different norm.

References

[1] J.A.E.E. van Nunen, J. Wessels, Markov decision processes with unbounded rewards. In this volume.

[2] J.A.E.E. van Nunen, Contracting Markov decision processes. Mathematical Centre Tract 71, Amsterdam 1976.

[3] J. Wessels, Stopping times and Markov programming. Transactions of the seventh Praque Conference on Information Theory, Statistical Decision Functions, Random Processes (including 1974 European Meeting of Statisticians). Academia, Prague (to appear).

[4] J;A.E.E. van Nunen, A set of successive approximation methods for discounted Markovian decision problems. Zeitschrift für Operations Res. 20 (1976) 203-208.

[5] J.A.E.E. van Nunen, Improved successive approximation methods for discounted Markov decision processes. p. 667-682 in A. Prékopa (ed.), Progress in Operation Research, Amsterdam, North-Holland Publ. Comp. 1976.

[6] J.A.E.E. van Nunen, J. Wessels, A principle for generating optimization procedures for discounted Markov decision processes. p. 683-695 in the same volume as [5].

[7] D. Reetz, Solution of a Markovian decision problem by overrelaxation. Z.f.O.R. 4 (1973) 29-32.

[8] D. Reetz, A decision exclusion algorithm for a class of Markovian decision processes. Zeitschrift für Operations Research, Vol. 20, 1976, 125-131.

[9] H. Schellhaas, Zur Extrapolation in Markoffschen Entscheidungsmodellen mit Diskontierung. Z.f.O.R. 18 (1974) 91-104.

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

[11] N.A.J. Hastings, J.A.E.E. van Nunen, The action elimination algorithm for Markov decision processes. In this volume.

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

[13] H.J. Kushner, A.J. Kleinman, Accelerated procedures for the solution of discrete Markov control problems. IEEE-Trans. on Aut. Contr. A.C. 16 (1971) 147-152.

[14] E.L. Porteus, Bounds and transformations for discounted finite Markov decision chains. Oper. Res. 23 (1975) 761-784.