Adv. Appl. Prob. **9**, 296–315 (1977)
Printed in Israel

A GENERAL MARKOV DECISION METHOD I: MODEL AND TECHNIQUES

G. DE LEVE, A. FEDERGRUEN AND
H. C. TIJMS, *Mathematisch Centrum, Amsterdam*

Abstract

This paper provides a new approach for solving a wide class of Markov decision problems including problems in which the space is general and the system can be continuously controlled. The optimality criterion is the long-run average cost per unit time. We decompose the decision processes into a common underlying stochastic process and a sequence of interventions so that the decision processes can be embedded upon a reduced set of states. Consequently, in the policy-iteration algorithm resulting from this approach the number of equations to be solved in any iteration step can be substantially reduced. Further, by its flexibility, this algorithm allows us to exploit any structure of the particular problem to be solved.

MARKOV DECISION PROBLEM; GENERAL STATE-SPACE; CONTINUOUS CONTROL; AVERAGE COST CRITERION; EMBEDDED DECISION PROCESSES; POLICY-ITERATION ALGORITHM

1. Introduction

This paper deals with a general Markov decision model introduced by De Leve [2]. In this model, which generalizes the familiar Markov decision models treated by Howard [10] and Jewell [12], the state-space is arbitrary, the system can be controlled at each point of time and the decision processes may be general Markov processes. The criterion is the long-run average cost per unit time. This paper treats the model studied in De Leve [2] under the simplifying assumption that any decision process has a fixed regeneration state, as is the case in almost any application. Under this assumption a self-contained exposition of the model will be given with proofs that have been considerably simplified. Emphasis will be put upon the presentation of a policy iteration method.

The approach we will follow is based on the following decomposition idea. Any decision process is considered as the result of a so-called natural process and interventions made in certain states of the system. The natural process could

Received in revised form 4 November 1976.

be considered as an underlying stochastic process which describes the evolution of the state of the system when the system is left uncontrolled, and interventions could be considered as those decisions that disturb the natural process. Also, the costs incurred in any decision process are decomposed into costs incurred in the natural process and immediate decision costs for taking interventions. It will appear that by this 'decomposition approach' we can in fact restrict ourselves to enbedded processes of the decision processes. This will have as main advantage that in the value-determination operation of the policy iteration method we need only solve a system of equations for an embedded set of states instead of a system of equations having the same order as the number of states. This may considerably reduce the number of equations to be actually solved. Another advantage of considering embedded processes is the fact that these processes have often desirable properties (e.g. recurrence properties) which are not satisfied by the original processes. It is also worth mentioning that the above approach allows for exploiting any structure of the particular problem to be solved, since this structure will be reflected in the natural process. However, as a consequence of the decomposition approach the policy iteration method needs another operation in addition to the familiar policy improvement routine. This operation is called the cutting operation, and actually involves the optimal stopping of a Markov process, as was stated in Weeda [16].

The policy iteration method we will derive is not a 'ready-made' technique, but its final form depends heavily on the structure of the particular problem to be solved. For each problem we have to specify the basic principles of our method. This flexibility which is inherent to our approach may result in a simple algorithm. Roughly speaking, we attack a particular problem to be solved as follows. First we perform a preparatory part in which we choose a state-space, a natural process and feasible decisions according to the following principles. The state-space must be such that at each point of time the state of the system can be represented by a point of the state-space. In each state of the system the set of feasible decisions contains 'interventions' which cause an instantaneous (possibly random) change of the state of the system or the 'null-decision' which leaves the natural process untouched, or both. The natural process and the feasible decisions must be chosen such that for any policy the corresponding decision process can be seen as a superimposition of the natural process and interventions prescribed by that policy. The choice of these elements will be determinative for the final form of the policy iteration method. Having made the above choices we can determine quantities which play similar roles to the one-step expected costs, the one-step expected transition times and the one-step transition probabilities in the familiar Markov decision models. We consider the average-cost criterion for the class of stationary policies. An iteration step of the policy iteration method for determining an optimal policy involves three operations. In the

value-determination operation we solve for the current policy a system of equations for the embedded set of states in which the policy prescribes an intervention. After we have computed the average cost and the relative values for the current policy, we next perform a policy improvement operation. By the design of our method this operation yields a new policy whose set of intervention states is at least as large as that of the previous policy. This implies the necessity for an additional operation which may cancel interventions in favour of the decision to leave the natural process untouched. This operation is called the cutting operation, and actually involves the optimal stopping of the natural process. Since in most applications the natural process is a Markov process having specific structure, the ultimate form of this cutting operation usually turns out to be very simple.

In Section 2 the basic elements of our model will be defined. The embedded decision processes are studied in Section 3 where we also derive a formula for the average cost of a given policy. This formula in itself may be very useful. In Section 4 we introduce the basic tools for the policy iteration method. This method will be further dicussed in Section 5. In the appendix we give several proofs. We should point out that in this paper we shall not discuss measurability questions. For a treatment of these questions we refer to De Leve [2]. Throughout this paper the words *set* and *function* serve as abbreviations for Borel set and Baire function.

In a subsequent paper [4] we shall discuss several applications of the approach given in the present paper.

2. The elements of the model

This section formulates the elements of the model. For any particular problem these elements have to be first specified before the actual solution of the problem can be started.

Element 1. *There is a state-space X such that at each point of time the state of the system can be described by a point in X, where X is a subset of a finite-dimensional Euclidean space.*

Element 2. *There is a stochastic process called the natural process. This process has X as state-space and could be considered as a process describing the evolution of the state of the system when the system is left uncontrolled. The natural process is a strong Markov process having stationary transition probabilities, and sample paths which are almost surely right-continuous and have a finite number of discontinuities in any finite time interval.*

We note that in applications the choice of the state-space and the natural process may involve the use of the supplementary variable technique. The natural process will be controlled by interventions.

Element 3. *For each state $x \in X$ there is a finite set $D(x)$ of feasible decisions in state x, where a distinction is made between null-decisions and interventions. A null-decision is a decision that does not disturb the natural process. An intervention is a decision that interrupts the natural process and causes an instantaneous (possibly random) change of the state of the system.*

We may assume that a transition caused by an intervention takes no time because at each point of time the state of the system is defined. In most applications the effect of an intervention is deterministic. The assumption of the finiteness of the sets of feasible decisions is not essential in this paper. The Elements 1–3 have to be chosen in such a way that the following element applies.

Element 4. *The states in which the null-decision is not feasible constitute a non-empty closed set A_0 (say) such that for each initial state, with probability* 1, *the natural process will eventually reach the set A_0. Further, with probability* 1, *any intervention in a state of A_0 causes an instantaneous transition to a state outside A_0.*

Element 5. *In the natural process there is incurred a cost at rate $c_1(x)$ when the system is in state x, and there is an immediate cost $c_2(x, y)$ at time t when the natural process is in state x at time t^- and is in state y at time t where $x \neq y$. There is incurred an immediate decision cost $c_3(x, d)$ when in state x the intervention $d \in D(x)$ is made. The functions c_1, c_2 and c_3 are non-negative.*

The non-negativity assumption in Element 5 is made only for convenience and may be considerably relaxed. In the next element we introduce the quantities $k(x; d)$ and $t(x; d)$. It will appear hereafter that in our model these quantities play the same role as the one-step expected costs and transition times in the semi-Markov decision model. The sets A_{01} and A_{02} introduced below are used only to define the functions $k(x; d)$ and $t(x; d)$ and may be freely chosen. We need the following notation which will be frequently used hereafter.

Let

$$X_0 = \{x \mid D(x) \text{ contains an intervention}\},$$

and for any $x \in X_0$ and intervention $d \in D(x)$, let

$T_{x,d}$ = the state into which the system is transferred instantaneously by the intervention d in state x.

Element 6. *Choose two non-empty closed sets $A_{01} \subseteq A_0$ and $A_{02} \subseteq A_0$ such that for each initial state, with probability* 1, *the natural process will eventually reach A_{0i} for $i = 1, 2$. Let $k_0(x) = 0$ for $x \in A_{01}$, and, for $x \notin A_{01}$, let $k_0(x)$ be the expected cost incurred up to and including the first epoch at which the system enters the set A_{01} when the system is subjected to the natural process and is in state x at*

epoch 0. *For any* $x \in X_0$ *and intervention* $d \in D(x)$, *let* $k_1(x;d) = c_3(x,d) + Ek_0(T_{x,d})$. *That is,* $k_1(x;d)$ *the expected cost incurred up to and including the first epoch at which the system enters* A_{01} *when at epoch* 0 *intervention* d *is made in state* x *and after this intervention the system is subjected to the natural process with the state resulting from this intervention as initial state. Similarly, let* $t_0(x) = 0$ *for* $x \in A_{02}$, *and, for* $x \notin A_{02}$, *let* $t_0(x)$ *be the expectation of the first epoch at which the system enters the set* A_{02} *when the system is subjected to the natural process and is in state* x *at epoch* 0. *For any* $x \in X_0$ *and intervention* $d \in D(x)$, *let* $t_1(x;d) = Et_0(T_{x,d})$. *It is assumed that* k_0, k_1, t_0 *and* t_1 *are finite functions. For any* $x \in X_0$ *and intervention* $d \in D(x)$, *let*

$$k(x;d) = k_1(x;d) - k_0(x) \quad \text{and} \quad t(x;d) = t_1(x;d) - t_0(x).$$

Hence $k(x;d)$ equals the immediate decision cost of intervention d in state x plus the expected cost incurred in the natural process until the first entry of this process into the set A_{01} starting from the state which is the immediate result from intervention d in state x minus the expected cost incurred in the natural process until it assumes for the first time a state of A_{01} starting from state x. Similarly, we can interprete $t(x;d)$.

The class of policies we will consider is denoted by Z and is described by the following element.

Element 7. *Any policy* $z \in Z$ *is a measurable function that adds to each state* $x \in X$ *a single decision* $z(x) \in D(x)$. *The states in which policy* $z \in Z$ *prescribes an intervention constitute a closed set* A_z *such that* $\Pr\{T_{x,z(x)} \in A_z\} = 0$ *for all* $x \in A_z$ *and* $\Pr\{T_{x,z(x)} \in A\}$ *is a Baire function of* $x \in A_z$ *for any set* A.

The process resulting from the control of the natural process by a policy $z \in Z$ is called the *decision process* corresponding to policy z. Between two successive interventions the behaviour of the decision process is described by the natural process. It is characteristic for our model to regard the decision process corresponding to policy z as a superimposition of the natural process and interventions made in the states of the embedded set A_z. Observe that $A_0 \subseteq A_z \subseteq X_0$. For any particular problem to be solved we have some freedom in choosing the Elements 1–3, provided that for any policy the superimposition of the natural process and the interventions prescribed by that policy agrees with the evolution of the system resulting from the specific control as executed by the decision-maker in reality. Since the final form of the policy iteration method will depend in a crucial way on the choice of the Elements 1–3, exploiting this freedom will turn out to enable considerable simplifications.

Remark 1. We may also consider a wider class of stationary policies z where the closedness of the intervention set A_z is not required provided that for

each initial state the entrance state of the natural process into the set A_z is well-defined. A similar remark applies to the closedness of the above sets A_0, A_{01} and A_{02}. Further, we may replace the assumption that, with probability 1, any intervention prescribed by policy z causes an instantaneous change to a state outside A_z by the weaker assumption that for each initial state $x \in A_z$, with probability 1, policy z transfers the system to a state outside A_z after a number of interventions uniformly bounded in x. The analysis below requires only minor modifications for this wider class of policies.

3. The embedded decision processes

In this section we derive a formula for the average cost of a policy in Z and introduce the system of equations to be solved in the value determination operation of the policy-iteration algorithm. Unless stated otherwise, we assume that a fixed policy z is used. In this section and the next one, we introduce assumptions A1–A6.

A1. *For any policy $z \in Z$ there are positive numbers δ_z and ε_z such that under policy z for each initial state $x \in A_z$ the probability that the time until the next return of the decision process to the set A_z exceeds δ_z is at least ε_z.*

This assumption implies that, with probability 1, the number of interventions is finite in any finite time interval.

We now introduce a discrete-time Markov process embedded in the decision process corresponding to policy z. Given that at epoch 0 the system is in state $x \in A_z$, define I_n as the state in which policy z prescribes for the nth time an intervention, $n = 0, 1, \cdots$ (at epoch 0 policy z prescribes for the 0th time an intervention). Using the strong Markov property of the natural process, it can be shown that $\{I_n\}$ is a discrete-time Markov process with state space A_z, cf. De Leve [2]. For $k = 0, 1, \cdots$, let $p^k(x, A, z) = \Pr\{I_k \in A \mid I_0 = x\}$ be the k-step transition probability function of the Markov chain $\{I_n\}$, where we write $p^1(x, A, z) = p(x, A, z)$. In the next assumption we assume that for each policy z the process $\{I_n\}$ has a fixed recurrent state.

A2. *For any policy $z \in Z$ there is some state s_z (say) such that $\Pr\{I_n = s_z$ for some $n \geqq 1 \mid I_0 = x\} = 1$ for all $x \in A_z$ and $E(N \mid I_0 = s_z) < \infty$ where $N = \inf\{n \geqq 1 \mid I_n = s_z\}$.*

Now, by a general result in Markov chain theory (see Theorem 9 in the appendix), the Markov chain $\{I_n\}$ has a unique stationary probability distribution $Q(\cdot, z)$ (say) such that

$$Q(A, z) = \lim_{n \to \infty} n^{-1} \sum_{k=0}^{n} p^k(x, A, z) \quad \text{for all } x \text{ and } A, \tag{1}$$

(2) $$Q(A, z) = \int_{A_z} p(y, A, z)\, Q\,(dy, z) \quad \text{for all } A.$$

A3. *For any policy* $z \in Z$,

(a) $$\int_{A_z} k_0(x) Q(dx, z) < \infty \quad \text{and} \quad \int_{A_z} t_0(x) Q(dx, z) < \infty.$$

(b) *For each initial state* $x \in X$ *holds that under policy* z *both the time until the first return of the decision process to state* s_z *and the cost incurred during this time have a finite expectation.*

For the decision process corresponding to policy z, let $Z(t)$ be the total cost incurred during $[0, t)$, $t > 0$. We shall now derive a formula for the average cost of policy z.

Theorem 1. *Suppose that* A1–A3 *hold. Then for each initial state,* $Z(t)/t$ *converges for* $t \to \infty$ *both in expectation and with probability* 1 *to*

(3) $$g(z) = \int_{A_z} k(x; z(x)) Q(dx, z) \Big/ \int_{A_z} t(x; z(x)) Q(dx, z).$$

Proof. For $n \geqq 0$, let T_n be the epoch at which policy z prescribes for the nth time an intervention, and let K_n be the decision cost of the nth intervention plus the other cost incurred in $(T_n, T_{n+1}]$. For any $x \in A_z$, let $\tau(x, z) = E(T_1 \mid I_0 = x)$ and let $\kappa(x, z) = E(K_0 \mid I_0 = x)$.

Consider first the case where the initial state is s_z. Following the proof of Theorem 7.5 in Ross [13] and using A1, A2 and A3(b), we get

(4) $$\lim_{t\to\infty} t^{-1} EZ(t) = \lim_{n\to\infty} n^{-1} \sum_{i=0}^{n} EK_i \Big/ \lim_{n\to\infty} n^{-1} \sum_{i=0}^{n} E(T_{i+1} - T_i),$$

where both the numerator and the denominator of the right-hand side of (4) are finite. Next, by using a general result in Markov chain theory (see Theorem 9 in the appendix), we get

(5) $$\lim_{t\to\infty} t^{-1} EZ(t) = \int_{A_z} \kappa(x, z) Q(dx, z) \Big/ \int_{A_z} \tau(x, z) Q(dx, z).$$

We note that the finiteness of both the numerator and the denominator of the right-hand side of (5) follows from the non-negativity of κ and τ, Proposition 17 on p. 231 in Royden [14] and the finiteness of both components of the ratio in (4). We now prove

(6) $$\int_{A_z} \kappa(x, z)\, Q\,(dx, z) = \int_{A_z} k(x; z(x))\, Q\,(dx, z)$$

(7) $$\int_{A_z} \tau(x, z)\, Q\,(dx, z) = \int_{A_z} t(x; z(x))\, Q\,(dx, z).$$

Observe that the right-hand side of (7) is positive since the function τ is positive. In Element 6 we have introduced the sets A_{01} and A_{02} and the functions k_0, k_1, t_0 and t_1. By $A_z \supseteq A_0$ we have $A_z \supseteq A_{0i}$ for $i = 1, 2$. Using this and the definitions of the functions k_0, k_1, t_0, t_1, κ and τ, it is easy to see

$$k_1(x; z(x)) = \kappa(x, z) + \int_{A_z} k_0(y) p(x, dy, z) \qquad \text{for all } x \in A_z,$$

$$t_1(x; z(x)) = \tau(x, z) + \int_{A_z} t_0(y) p(x, dy, z) \qquad \text{for all } x \in A_z.$$

Now, integrate both sides of each of these equalities with respect to $Q(\cdot, z)$. Using the non-negativity of the functions involved, Part (a) of A3, relation (2) and the finiteness of both components of the ratio in (5), we get after an interchange of the order of integration the desired equations (6) and (7). Together (5)–(7) prove that $EZ(t)/t$ converges to $g(z)$ as $t \to \infty$. However, using Part (b) of A3, it is easy to verify that $\lim_{t\to\infty} EZ(t)/t$ is independent of the initial state. Moreover, by Theorem 3.16 in Ross [13], we have for each initial state, with probability 1, $\lim_{t\to\infty} Z(t)/t$ equals $\lim_{t\to\infty} EZ(t)/t$. This ends the proof.

The quantity $g(z)$ represents the *long-run average (expected) cost per unit time* when policy z is used. This quantity is independent of the initial state. A policy $z^* \in Z$ is called *optimal* when $g(z^*) \leqq g(z)$ for all $z \in Z$.

The average cost $g(z)$ can also be found by solving an embedded system of equations. In solving this system we obtain in addition a function that will be used to improve policy z.

A4. *For any policy* $z \in Z$,
(a) $E(N \mid I_0 = x)$ *is bounded in* $x \in A_z$ *where* $N = \inf\{n \geqq 1 \mid I_n = s_z\}$,
(b) $k(x; z(x))$ *and* $t(x; z(x))$ *are bounded functions of* $x \in A_z$.

Consider now the following system of functional equations for the states of A_z,

$$v(x) = k(x; z(x)) - gt(x; z(x)) + Ev(I_1 \mid I_0 = x), \qquad x \in A_z. \tag{8}$$

For any bounded solution $\{g, v(x) \mid x \in A_z\}$ to (8), define

$$v(x) = Ev(S[x, A_z]) \quad \text{for } x \notin A_z, \tag{9}$$

where for any $x \in X$ and closed set $A \supseteq A_0$ we define

$S[x, A]$ = the first state in the set A taken on by the natural process starting from state x.

Observe that by $A \supseteq A_0$ and Element 4, the random variable $S[x, A]$ is well-defined. Further, observe that $S[x, A] = x$ for $x \in A$.

The next theorem, which is related to Theorem 1 in Derman and Veinott [6], shows that a bounded solution to (8) exists.

Theorem 2. *Suppose that* A1–A4 *hold. Then*

(a) *Let* $g = g(z)$ *and, for* $x \in A_z$, *let*

$$w(x) = \sum_{n=0}^{\infty} \int_{A_z} \{k(y; z(y)) - gt(y; z(y))\} \hat{p}^n(x, dy, z), \tag{10}$$

where $\hat{p}^0(x, A, z) = 1$ *for* $x \in A$, $\hat{p}^0(x, A, z) = 0$ *for* $x \notin A$, *and for* $n \geqq 1$,

$$\hat{p}^n(x, A, z) = \Pr\{I_n \in A,\ I_k \neq s_z \quad \text{for} \quad 1 \leqq k \leqq n \mid I_0 = x\}.$$

Then $\{g, w(x) \mid x \in A_z\}$ *is a bounded solution to* (8) *with* $w(s_z) = 0$.

(b) *For any bounded solution* $\{g, v(x)\}$ *to* (8) $g = g(z)$ *holds.*

(c) *For any two bounded solutions* $\{g, v_1(x)\}$ *and* $\{g, v_2(x)\}$ *to* (8) *there is a constant* c *such that* $v_1(x) - v_2(x) = c$ *for all* $x \in A_z$.

(d) *Let* y *be an arbitrary state in* X, *then* (8) *and* (9) *together have a unique bounded solution with* $v(y) = 0$.

Proof. (a) By $EN = \Sigma_{n=0}^{\infty} \Pr\{N > n\}$, we have

$$E(N \mid I_0 = x) = \Sigma_{n=0}^{\infty} \hat{p}^n(x, A_z, z) \qquad \text{for } x \in A_z.$$

Using this and A4, we get that $w(x)$ is bounded. From (3) and the relation $Q(A, z) = \Sigma_0^{\infty} \hat{p}^n(s_z, A, z)/E(N \mid I_0 = s_z)$ (see (24) in the appendix), we get $w(s_z) = 0$. Using this, A4 and the relation

$$\hat{p}^n(x, A, z) = \int_{A_z \setminus \{s_z\}} \hat{p}^{n-1}(y, A, z) p(x, dy, z) \qquad \text{for } n \geqq 1,$$

we next find that $\{g(z), w(x) \mid x \in A_z\}$ satisfies (8).

(b) Integrating both sides of (8) with respect to $Q(\cdot, z)$, and using the relations (2) and (3), we get (b).

(c) Using Part (b), we have $v_1(x) - v_2(x) = \int \{v_1(y) - v_2(y)\} p(x, dy, z)$ for $x \in A_z$. Iterate this equality n times and average over n. Letting $n \to \infty$ and using (1), we get $v_1(x) - v_2(x) = \int \{v_1(y) - v_2(y)\} Q(dy, z)$ for $x \in A_z$ which proves (c).

(d) This assertion follows from (a)–(c) and the fact that $\{g, v(x) + \gamma\}$ satisfies (8)–(9) for any constant γ when $\{g, v(x)\}$ is a solution to (8)–(9).

Remark 2. In this remark we first give some relations which may be useful in solving (8)–(9) in applications. Let $\{g, v(x)\}$ be any solution to (8)–(9). It immediately follows from (8)–(9) that

$$v(x) = k(x; z(x)) - gt(x; z(x)) + Ev(T_{x,z(x)}) \qquad \text{for } x \in A_z. \tag{11}$$

Further, let V be any closed set with $V \supseteq A_z$. Then, by (9) and the theorem of conditional expectation,

(12) $$v(x) = Ev(S[x, V]) \quad \text{for } x \notin A_z.$$

We further note that in fact we need only solve (8) in order to obtain a solution to (8)–(9). The dimension of the system of equations (8) is equal to the dimension of the embedded set A_z. However, the dimension of A_z is determined by the choice of the natural process because an intervention is a decision which interrupts the natural process. Therefore, to keep the number of equations to be solved as small as possible it may be advantageous to make 'obvious optimal decisions' part of the natural process. Finally, we note that in applications in which the effects of the interventions are deterministic we may often solve (8)–(9) by solving a system of equations for the states $T_{x,z(x)}$, $x \in A_z$. In particular for structured policies z the state $T_{x,z(x)}$ may be the same for different states $x \in A_z$ which has a consequence that the dimension of the latter system of equations may be even considerably smaller than the dimension of the set A_z. This observation also underlines the advantage of choosing the state-space such that at each point of time the state of the system is well-defined.

4. Basic tools for the solution techniques

This section discusses the basic tools for the solution techniques. We fix a policy $z_1 \in Z$ and a bounded solution $\{g(z_1), v(z_1; x)\}$ to (8)–(9) with $z = z_1$. We now define for any $x \in X$ and $d \in D(x)$,

(13) $$v(d.\, z_1; x) = \begin{cases} v(z_1; x) & \text{for } d = \text{null-decision}, \\ k(x; d) - g(z_1)t(x; d) + Ev(z_1; T_{x,d}) & \text{otherwise.} \end{cases}$$

Further, for any policy $z \in Z$, define

(14) $$v([z]z_1; x) = \begin{cases} v(z(x).\, z_1; x) & \text{for } x \in A_z, \\ Ev([z]z_1; S[x, A_z]) & \text{for } x \notin A_z. \end{cases}$$

Note that, by (11) and (13), $v(z_1(x).\, z_1; x) = v(z_1; x)$ for all x, so, by (9) and (14),

(15) $$v([z_1]z_1; x) = v(z_1; x) \qquad \text{for all } x \in X.$$

We now state the following main theorem which will be proved in the appendix.

Theorem 3. *Suppose that* A1–A4 *hold. Let policy* $z \in Z$ *be such that* $v([z]z_1; x) \leqq v(z_1; x)$ *for all* $x \in X_0$. *Then* $g(z) \leqq g(z_1)$. *The assertion remains true when both inequality signs are reversed.*

This theorem implies that policy z_1 is optimal when

(16) $$v(z_1; x) = \min_{z \in Z} v([z]z_1; x) \qquad \text{for all } x \in X_0.$$

This relation may provide a direct approach for determining an optimal policy. See [3] and [4] for applications. However, in most cases an iterative approach will be used. When we want to improve policy z_1, relation (16) suggests we could look for a policy $z_2 \in Z$ such that

$$v([z_2]z_1; x) = \min_{z \in Z} v([z]z_1; x) \qquad \text{for all } x \in X_0. \tag{17}$$

Then by (15) and (17) and Theorem 3, $g(z_2) \leqq g(z_1)$. We shall now prove that a policy z_2 satisfying can be found by performing two operations. To do this, we need the following concept. For any policy $z \in Z$ and closed set $A \supseteq A_0$, let

$$v(A.[z]z_1; x) = Ev([z]z_1; S[x, A]), \qquad x \in X. \tag{18}$$

We write $v(A.[z_1]z_1; x) = v(A.z_1; x)$. By (15) and (18),

$$v(A.z_1; x) = Ev(z_1; S[x, A]), \qquad x \in X. \tag{19}$$

It may be helpful to interpret $v(A.[z]z_1; x)$ as the expected stopping cost for the natural process starting from state x when this process must be stopped at the states of the set A and there is a cost of $v([z]z_1; y)$ for stopping at state y. The next theorem which will be proved in the appendix shows that the stopping principle given by (18) enables us to generate improved policies.

Theorem 4. *Suppose that* A1–A4 *hold. Let* $z \in Z$ *be such that* $A_z \supseteq A_{z_1}$ *and* $v([z]z_1; x) \leqq v(z_1; x)$ *for all* $x \in A_z$. *Let* A *be any closed set with* $A_0 \subseteq A \subseteq A_z$ *such that* $v(A.[z]z_1; x) \leqq v([z]z_1; x)$ *for all* $x \in A_z$. *Suppose that policy* $z_A \in Z$ *where* $z_A(x) = z(x)$ *for* $x \in A$, *and* $z_A(x) =$ *null-decision, otherwise. Then* $g(z_A) \leqq g(z_1)$.

This theorem states that policy z_A is at least as good as policy z_1 if for the natural process the set A is a stopping set at least as good as the set A_z for each initial state of A_z when there is a cost of $v([z]z_1; y)$ for stopping at state y. An important case of Theorem 4 arises when $z = z_1$.

The next lemma which follows from Lemma 10 to be proved in the appendix shows that $A_1 \cap A_2$ has the properties of the set A in Theorem 4 if both A_1 and A_2 do. For this useful lemma, we need the following assumption.

A5. *For any closed set* A *such that* $A_0 \subseteq A \subseteq X_0$ *and for each initial state, the number of times where the natural process enters the set* A *before it enters the set* A_0 *is finite with probability* 1.

Lemma 5. *Suppose that* A1–A5 *hold. Let policy* z *be as in Theorem* 4 *and let the sets* A_1 *and* A_2 *be as the set* A *in Theorem* 4. *Then, for the natural process the set* $A_1 \cap A_2$ *is as stopping set at least as good as each of the sets* A_1 *and* A_2 *for each initial state of* A_z *when there is a cost of* $v([z]z_1; y)$ *for stopping at state* y.

We now have available the tools for determining a policy z_2 satisfying (17). To do this, we first perform a policy improvement operation in which we add to each state $x \in X_0$ a decision $d \in D(x)$ for which $v(d.z_1; x)$ is minimal where we choose $d = z_1(x)$ when this decision minimizes $v(d.z_1; x)$. In this way we obtain a policy z_1'. It is assumed that $z_1' \in Z$. By the construction of z_1' and the fact that $v(d.z_1; x)$ assumes the same value for both $d = z_1(x)$ and $d =$ null-decision, we have

$$A_{z_1'} \supseteq A_{z_1}. \tag{20}$$

By taking $z = z_1'$ and $A = A_z$ in Theorem 4, we have $g(z_1') \leqq g(z_1)$. Although we obtain an improved policy, it will be clear from (20) that we need a second operation which may replace interventions prescribed by policy z_1' by null-decisions. This so-called cutting operation will yield the desired policy z_2. Therefore we need another assumption.

A6. (a) *There is a non-empty class $\mathcal{R}$ of closed sets A with $A_0 \subseteq A \subseteq X_0$ such that, for all $x \in X_0$, $v(A.[z_1']z_1; x) \leqq v(B.[z_1']z_1; x)$ for any closed set B with $A_0 \subseteq B \subseteq X_0$.*

(b) *The intersection of all sets belonging to $\mathcal{R}$ belongs also to $\mathcal{R}$.*

Observe that A6(a) requires that there is a set A which is an optimal stopping set for the natural process for each initial state of X_0 when the natural process must be stopped at the states of A_0, may be stopped at the states of X_0, and must be continued outside X_0 and there is a cost of $v([z_1']z_1; y)$ for stopping at state y. By a well-known result in the theory of optimal stopping (cf. Chapter 8 in Derman [7]), A6(a) holds when $X_0 \backslash A_0$ is finite. Further, we note that, by Lemma 10 in the appendix, A6(b) holds when $\mathcal{R}$ is finite.

In the appendix we shall prove the following result.

Lemma 6. *Suppose that* A1–A4 *and* A6(a) *hold. Let $A \in \mathcal{R}$ be such that $A \subseteq A_{z_1}$ and that policy $z_A \in Z$ where $z_A(x) = z_1'(x)$ for $x \in A$, and $z_A(x) =$ null-decision, otherwise. Then $z_2 = z_A$ satisfies* (17).

The next theorem which will be proved in the appendix shows that a set $A \in \mathcal{R}$ such that $A \subseteq A_{z_1'}$ can be constructed.

Theorem 7. *Suppose that* A1–A6 *hold. Let R^* be the intersection of all sets belonging to $\mathcal{R}$. Then*

(a) *$R^* \in \mathcal{R}$ and $R^* \subseteq A_{z_1}$,*

(b) *R^* is the smallest closed set A with $A_0 \subseteq A \subseteq A_{z_1'}$ such that, for any closed set B with $A_0 \subseteq B \subseteq A_{z_1}$,*

$$v(A.[z_1']z_1; x) \leqq v(B.[z_1']z_1; x) \quad \text{for all } x \in A_{z_1}. \tag{21}$$

(c) *Let A be any closed set with $A_0 \subseteq A \subseteq A_{z_1'}$ such that* (21) *holds for all closed sets B with $A_0 \subseteq B \subseteq A_{z_1'}$. Suppose that policy $z_2 \in Z$, where $z_2(x) = z_1'(x)$ for $x \in A$, and $z_2(x) =$ null-decision, otherwise. Then z_2 satisfies* (17).

Since policy z_2 satisfying (17) is optimal when $z_2 = z_1$ (see (15)–(16)), Theorem 7 has the following corollary.

Theorem 8. *Suppose that* A1–A5 *and* A6(a) *with $z_1' = z_1$ hold. Then policy z_1 is optimal when*

(a) *$v(d.z_1; x) \geqq v(z_1; x)$ for all $x \in X_0$ and $d \in D(x)$,*

(b) *$v(B.z_1; x) \geqq v(z_1; x)$ for all $x \in A_{z_1}$ and all closed sets B with $A_0 \subseteq B \subseteq A_{z_1}$.*

5. Policy iteration algorithms

In this section we give a policy iteration algorithm and a modification of this method. As already stated, in solving any particular problem we have first to specify the Elements 1–6 for this problem. We now give a policy iteration algorithm which generates a sequence $\{z_n\}$ of policies such that (17) applies for all n when z_1 and z_2 are replaced by z_n and z_{n+1}.

Policy iteration algorithm

Let z_n be the policy obtained at the end of the $(n-1)$th iteration step (the first step is started with an arbitrary policy $z_1 \in Z$). The nth step of the policy iteration algorithm proceeds as follows.

(a) *Value-determination operation.* Determine a bounded solution $\{g(z_n), v(z_n; x)\}$ to (8) and (9) with $z = z_n$.

(b) *Policy improvement operation.* Construct policy z_n' by adding to each state $x \in X_0$ a decision $d \in D(x)$ for which

$$v(d.z_n; x) = \begin{cases} v(z_n; x) & \text{for } d = \text{null-decision}, \\ k(x; d) - g(z_n)t(x; d) + Ev(z_n; T_{x,d}) & \text{otherwise}, \end{cases}$$

is minimal, where $z_n'(x) = z_n(x)$ is chosen when $z_n(x)$ is a minimizing decision.

(c) *Cutting operation.* Determine a closed set A with $A_0 \subseteq A \subseteq A_{z_n'}$ such that A is an optimal stopping set for the natural process for each initial state $x \in A_{z_n'}$ when the natural process must be stopped at the states of A_0, may be stopped at the states of $A_{z_n'}$ and must be continued outside $A_{z_n'}$ and there is a cost of $v(z_n'(y).z_n; y)$ for stopping at state y. Define policy z_{n+1} by $z_{n+1}(x) = z_n'(x)$ for $x \in A$, and $z_{n+1}(x) =$ null-decision, otherwise.

This policy iteration method generates a sequence $\{z_n\}$ of policies where it is assumed that $z_n, z_n' \in Z$ for all $n \geqq 1$. If follows from the Theorems 3 and 7 that

$g(z_{n+1}) \leqq g(z_n)$ for all n. Further, policy z_k is optimal when $z_{k+1} = z_k$. Under A1–A6 and the additional assumption that both state s_z in A3 and the bounds in A4 can be taken independently of $z \in Z$, it can be shown that $\lim_{n\to\infty} g(z_n) = \inf_{z\in Z} g(z)$, see [5] where such a convergence result has been also established under a recurrency condition which does not assume the existence of fixed regeneration states for the decision processes.

We shall now discuss a modified policy iteration method which is based on Theorem 4. Therefore we first note that a policy $f \in Z$ may be improved to a policy $f' \in Z$ (say) by adding to each state $x \in X_0$ any decision d such that $v(d.f; x) \leqq v(f; x)$ provided that we never choose d = null-decision in state x when $f(x)$ is an intervention. Observe that this can always be done since $v(f(x).f; x) = v(f; x)$. Then, we have $A_{f'} \supseteq A_f$ and $v([f']f; x) \leqq v(f; x)$ for all $x \in A_{f'}$.

We now state the following algorithm.

Modified policy iteration algorithm

Let policy $f \in Z$ be given.

(a) Determine a bounded solution $\{g(f), v(f; x)\}$ to (8)–(9) with $z = f$.

(b) Construct a policy $f' \in Z$ by adding to each state $x \in X_0$ a decision $d \in D(x)$ with $v(d.f; x) \leqq v(f; x)$ such that d is an intervention when $f(x)$ is an intervention.

(c) Construct policy $f'' \in Z$ by taking $f''(x) = f'(x)$ for $x \in A$ and $f''(x)$ = null-decision, otherwise, where A is any closed set with $A_0 \subseteq A \subseteq A_{f'}$ such that for the natural process the set A is as stopping set at least as good as the set $A_{f'}$ for each initial state $x \in A_{f'}$ when there is a cost of $v(f'(y).f; y)$ for stopping at state y.

Remark 3. Taking $z = z_1 = f'$ in Theorem 4 and taking into account the relations (15) and (19), we have that instead of step (c) of the above method the following step may be applied.

(c′) Determine a bounded solution $\{g(f'), v(f'; x)\}$ to (8) with $z = f'$. Construct policy f'' as in the above step (c) by now taking $v(f'; y)$ as cost for stopping at state y.

The modified policy iteration method may be useful particularly to generate a sequence of policies having each a prescribed structure, cf. [4]. Moreover, from a computational point of view the modified method may be more attractive than the policy iteration method first stated, cf. Weeda [15], [16] where in addition for the case of a finite X_0 conditions for the construction of f' and f'' have been derived under which the modified method converges to an optimal policy after a finite number of iterations. In general we note that Theorem 8 may be applied to check whether a given policy is optimal among the class Z of policies.

Remark 4. To generate a set A as in step (c) or (c′) of the modified method the following procedure may be useful when $A_{f'}$ is countable. For a properly chosen sequence of points $x \in A_{f'}$, we may take the set A as the intersection of all those sets $A_{f'}\backslash\{x\}$ such that for the natural process starting from state x the set $A_{f'}\backslash\{x\}$ is as stopping set at least as good as the set $A_{f'}$, see Lemma 5. Note that for each x this involves only the verification of a single inequality.

6. Appendix

We now give some results for discrete-time Markov processes with a general state-space and we give the proofs of the theorems and lemmas of Section 4.

Consider a Markov chain $X_0, X_1, X_2, \cdots$ with stationary transition probability function $p(.,.)$ on $(S, \mathcal{B})$ where the state-space is a Borel set of a finite-dimensional Euclidean space and $\mathcal{B}$ is the class of all Borel sets in S. For any $n \geqq 0$, let $p^n(.,.)$ be the n-step transition probability function of the Markov chain. That is, $p^n(x, A) = \Pr\{X_n \in A \mid X_0 = x\}$. We assume that there is some state s (say) such that

$$\Pr\{X_n = s \text{ for some } n \geqq 1 \mid X_0 = x\} = 1 \qquad \text{for all } x \in S, \tag{22}$$

$$E(N \mid X_0 = s) < \infty \text{ where } N = \inf\{n \geqq 1 \mid X_n = s\}. \tag{23}$$

Let $\hat{p}^0(x, A) = 1$ for $x \in A$, let $\hat{p}^0(x, A) = 0$ for $x \notin A$, and let

$$\hat{p}^n(x, A) = \Pr\{X_n \in A,\ X_k \neq s \text{ for } 1 \leqq k \leqq n \mid X_0 = x\} \text{ for } n \geqq 1.$$

For any set $A \in \mathcal{B}$, define

$$Q(A) = \sum_{n=0}^{\infty} \hat{p}^n(s, A) / E(N \mid X_0 = s). \tag{24}$$

Observe that, by $EN = \Sigma_0^\infty \Pr\{N > n\}$,

$$E(N \mid X_0 = s) = \sum_{n=0}^{\infty} \hat{p}^n(s, S), \tag{25}$$

so, $Q(\cdot)$ is a probability distribution. We note that $Q(A)$ can be interpreted as the ratio of the expected number of visits of the Markov chain to the set A before returning to state s and the expected number of transitions needed to return to state s starting from state s.

Theorem 9. *For any* $A \in \mathcal{B}$,

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n} p^k(x, A) = Q(A) \qquad \textit{for all } x \in S \tag{26}$$

$$Q(A) = \int_S p(x, A) Q(dx). \tag{27}$$

Further, Q is the unique stationary probability distribution of the Markov chain $\{X_n\}$. Also, when $X_0 = s$,

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n} Ef(X_k) = \int_S f(x)Q(dx) \tag{28}$$

for any Baire function f such that $\int |f(x)|\, Q(dx)$ is finite.

Proof. For any $x \in S$, let $f_0(x) = 0$, and let $f_n(x) = \Pr\{N = n \mid X_0 = x\}$, $n \geqq 1$. By (22), $\Sigma_0^\infty f_n(x) = 1$ for all x. Clearly, for any x and A (cf. p. 365 in Feller [9]),

$$p^n(x, A) = \hat{p}^n(x, A) + \sum_{k=0}^{n} p^{n-k}(s, A) f_k(x) \qquad \text{for } n \geqq 0. \tag{29}$$

For $x = s$ this relation is a renewal equation. By (23) and (25), both $\Sigma n f_n(s)$ and $\Sigma \hat{p}^n(s, A)$ are finite. Now, by applying the key renewal theorem (see p. 292 in Feller [8]), for any $A \in \mathcal{B}$,

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n} p^k(s, A) = \sum_{n=0}^{\infty} \hat{p}^n(s, A) \Big/ \sum_{n=0}^{\infty} n f_n(s) = Q(A). \tag{30}$$

Since $\Sigma_0^\infty f_n(x) = 1$ and $\hat{p}^n(x, A) \to 0$ as $n \to \infty$ for all x and A, relation (26) now follows from (29) and (30). Using (26) it is easy to verify that Q satisfies the steady state equation (27) (cf. pp. 133–134 in Breiman [1]). Since the Markov chain $\{X_n\}$ has no two disjoint closed sets, Q is the unique probability distribution satisfying (27), see Theorem 7.16 in Breiman [1]. To prove (28), let m be a finite measure on $(S, \mathcal{B})$ such that $m(A) > 0$ if and only if $s \in A$. Then, by (22), $m(A) > 0$ implies $\Pr\{X_n \in A \text{ for some } n \geqq 1 \mid X_0 = x\} = 1$ for all $x \in S$. Consequently, the Markov chain $\{X_n\}$ satisfies the recurrence condition of Harris (cf. pp. 206–207 in Jain [11]). Relation (28) now follows from Theorem 3.3 in Jain [11].

Before giving the proofs of the theorems and lemmas of Section 4, we state some relations which will be frequently used in these proofs. Let V and W be any closed sets with $V \supseteq A_{z_1}$ and $W \supseteq A_z$ for the policies z_1 and z. Then, using the definitions (9) and (14) and the theorem of conditional expectation, we have for all $x \in X$,

$$v(z_1; x) = Ev(z_1; S[x, V]) \quad \text{and} \quad v([z]z_1; x) = Ev([z]z_1; S[x, W]). \tag{31}$$

Similarly, the following relation applies to (18). Let V be any closed set with $V \supseteq A$, then for all $x \in X$

$$v(A.[z]z_1; x) = Ev(A.[z]z_1; S[x, V]). \tag{32}$$

Proof of Theorem 3. Since $v([z]z_1; x) \leqq v(z_1; x)$ for $x \in X_0$, it follows from (31) with $V = W = X_0$ that

$$v([z]z_1; x) = Ev([z]z_1; S[x, X_0]) \leqq Ev(z_1; S[x, X_0]) = v(z_1; x), \qquad x \in X.$$

Hence, for all $x \in A_z$,

$$(33) \qquad k(x; z(x)) - g(z_1)t(x; z(x)) + Ev([z]z_1; T_{x,z(x)}) \leqq k(x; z(x)) - g(z_1)t(x; z(x)) + Ev(z_1; T_{x,z(x)}).$$

By (13) and (14), the right side of (33) equals $v([z]z_1; x)$. We have by (14) that

$$(34a) \qquad Ev([z]z_1; T_{x,z(x)}) = \int_{A_z} v([z]z_1; y)p(x, dy, z) \qquad \text{for } x \in A_z.$$

Hence, by (33), for all $x \in A_z$,

$$(34b) \qquad k(x; z(x)) - g(z_1)t(x; z(x)) + \int_{A_z} v([z]z_1; y)p(x, dy, z) \leqq v([z]z_1; x).$$

Now, integrate both sides of (34a) with respect to $Q(.\,, z)$. Using the boundedness of the functions k, t and v, and using the relations (2) and (3) we get after an interchange of the order of integration $g(z) \leqq g(z_1)$. Clearly, this proof carries over when the inequality signs are reversed.

Proof of Theorem 4. We first prove $v(A.[z]z_1; x) = v([z_A]z_1; x)$ for all $x \in X$. For $x \in A$ this equality follows immediately from the relation $v(A.[z]z_1; x) = v([z]z_1; x)$, (14) and the fact that $z_A(x)$ is an intervention which equals $z(x)$. For $x \notin A$ we next have by (32) with $V = A$ and (14),

$$v(A.[z]z_1; x) = Ev(A.[z]z_1; S[x, A]) = Ev([z_A]z_1; S[x, A]) = v([z_A]z_1; x).$$

Next we prove $v(A.[z]z_1; x) \leqq v(z_1; x)$ for all $x \in X$. By the conditions of the Theorem, this inequality holds for $x \in A_z$. Since $A \subseteq A_z$ and $A_{z_1} \subseteq A_z$ it follows from (32) and (31) with $V = A_z$ that, for $x \notin A_z$,

$$v(A.[z]z_1; x) = Ev(A.[z]z_1; S[x, A_z]) \leqq Ev(z_1; S[x, A_z]) = v(z_1; x).$$

Together the above relations yield $v([z_A]z_1; x) \leqq v(z_1; x)$ for all $x \in X$, so, by Theorem 3, $g(z_A) \leqq g(z_1)$.

Lemma 10. *Suppose that* A1–A5 *hold. Let* $u(x)$ *be a bounded function on* X. *Let* A_1 *and* A_2 *be closed sets with* $A_0 \subseteq A_i \subseteq X_0$ *for* $i = 1, 2$. *For* $i = 1, 2$ *and* $x \in X$, *let* $v_i(x) = Eu(S[x, A_i])$, *and let* $v(x) = Eu(S[x, A_1 \cap A_2])$. *Suppose that* $v_i(x) \leqq u(x)$ *for* $i = 1, 2$ *and all* $x \in A_1 \cup A_2$. *Then* $v(x) \leqq v_i(x)$ *for* $i = 1, 2$ *and all* $x \in X$.

Proof of Lemma 10. For reasons of symmetry it suffices to prove $v_1(x) \geqq v(x)$ for all $x \in X$. Clearly, this inequality holds with the equality sign for $x \in A_1 \cap A_2$. Let $P(B \mid x, A) = \Pr\{S(x, A) \in B\}$. Now fix $x \in A_1^c$ where $A^c = X \backslash A$. Using the fact that $u(y) \geqq v_2(y)$ for all $y \in A_1$, we get

$$v_1(x) = \int_{A_2} u(y_1)P(dy_1 \mid x, A_1) + \int_{A_2^c} u(y_1)P(dy_1 \mid x, A_1)$$

$$\geqq \int_{A_2} u(y_1)P(dy_1 \mid x, A_1) + \int_{A_2^c} P(dy_1 \mid x, A_1)\Big\{\int_{A_1} u(y_2)P(dy_2 \mid y_1, A_2)$$

$$+ \int_{A_1^c} u(y_2)P(dy_2 \mid y_1, A_2)\Big\}$$

Using the fact that $u(y) \geqq v_1(y)$ for all $y \in A_2$, we next get

$$\int_{A_2^c} P(dy \mid x, A_1) \int_{A_1^c} u(y_2)P(dy_2 \mid y_1, A_2)$$

$$\geqq \int_{A_2^c} P(dy_1 \mid x, A_1) \int_{A_1^c} P(dy_2 \mid y_1, A_2)\Big\{\int_{A_2} u(y_3)P(dy_3 \mid y_2, A_1)$$

$$+ \int_{A_2^c} u(y_3)P(dy_3 \mid y_2, A_1)\Big\}.$$

Continuing in this way yields for $n = 2, 3, \cdots$

$$v_1(x) \geqq \int_{A_2} u(y_1)P(dy_1 \mid x, A_1) + \sum_{k=1}^{n-1} \int_{B_1^c} P(dy_1 \mid x, B_0)$$

$$\cdots \int_{B_k^c} P(dy_k \mid y_{k-1}, B_{k-1}) \int_{B_{k+1}} u(y_{k+1})P(dy_{k+1} \mid y_k, B_k) + c_n,$$

where

$$c_n = \int_{B_1^c} P(dy_1 \mid x, B_0) \cdots \int_{B_n^c} u(y_n)P(dy_n \mid y_{n-1}, B_{n-1}),$$

$$B_{2k} = A_1, \quad \text{and} \quad B_{2k+1} = A_2 \qquad \text{for } k = 0, 1, \cdots.$$

By A5 and the boundedness of $u(\cdot)$, $\lim_{n\to\infty} c_n = 0$. Further, for any set B,

$$P(B \mid x, A_1 \cap A_2) = P(B \cap A_2 \mid x, A_1)$$

$$+ \int_{A_2^c} P(dy_1 \mid x, A_1)P(B \cap A_1 \mid y_1, A_2)$$

$$+ \int_{A_2^c} P(dy_1 \mid x, A_1) \int_{A_1^c} P(dy_2 \mid y_1, A_2)P(B \cap A_2 \mid y_2, A_1) + \cdots.$$

Using these relations we have $\int u(y)P(dy \mid x, A_1 \cap A_2)$ equals the limit of the right side of the latter inequality as $n \to \infty$. Hence $v_1(x) \geqq v(x)$ for all $x \in A_1^c$. For reasons of symmetry, $v_2(x) \geqq v(x)$ for all $x \in A_2^c$. From this we get $v_1(x) = u(x) \geqq v_2(x) \geqq v(x)$ for all $x \in A_1 \setminus (A_1 \cap A_2)$. We now have proved $v_1(x) \geqq v(x)$ for all $x \in X$, which ends the proof.

Proof of Lemma 6. By the first part of the proof of Theorem 4,

$$v(A.[z_1']z_1; x) = v([z_2]z_1; x) \qquad \text{for all } x \in X. \tag{35}$$

We shall next prove that, for all $x \in X$,

$$v(z_1; x) \geqq v([z_1']z_1; x). \tag{36}$$

Clearly, by (13)–(15) and the construction of z_1', this inequality holds for $x \in A_{z_1}$. Next it follows from (31) with $V = A_{z_1}$ and (14) that (36) holds for all $x \in X$. By the construction of z_1' we have $v(z(x).z_1; x) \geqq v(z_1'(x).z_1; x)$ for all $x \in X$ and $z \in Z$. Distinguishing between $x \in A_{z_1}$ and $x \notin A_{z_1}$ it now follows from the latter inequality, (36) and the definitions (13) and (14) that, for any policy $z \in Z$,

$$v([z]z_1; x) \geqq v([z_1']z_1; x) \qquad \text{for all } x \in A_z. \tag{37}$$

By (18), (37) and (14), for all $z \in Z$ and $x \in X$,

$$v(A_z.[z_1']z_1; x) \leqq Ev([z]z_1; S[x, A_z]) = v([z]z_1; x). \tag{38}$$

Assume now to the contrary that $v([z_0]z_1; x_0) < v([z_2]z_1; x_0)$ for some $z_0 \in Z$ and $x_0 \in X_0$. Together this inequality, (35) and (38) contradict the inequality in A6(a). Hence z_2 satisfies (17).

Proof of Theorem 7. (a) Let $\mathcal{K}$ be the class of all closed sets A with $A_0 \subseteq A \subseteq X_0$ such that $v(A.[z_1']z_1; x) \leqq v([z_1']z_1; x)$ for all $x \in X$. Since $v(A_z.[z]z_1; x) = v([z]z_1; x)$ for all $x \in X$ and $z \in Z$, we have

$$A_{z_1} \in \mathcal{K} \quad \text{and} \quad \mathcal{R} \subseteq \mathcal{K}. \tag{39}$$

Further, by taking $u(x) = v([z_1']z_1; x)$ in Lemma 10 and using $\mathcal{R} \subseteq \mathcal{K}$, we easily get

$$A_1 \cap A_2 \in \mathcal{R} \quad \text{when} \quad A_1 \in \mathcal{R} \quad \text{and} \quad A_2 \in \mathcal{K}. \tag{40}$$

Denote by K^* the intersection of all sets belonging to $\mathcal{K}$. We shall now prove $K^* = R^*$ which implies part (a) since $A_{z_1} \in \mathcal{K}$. Clearly, by $\mathcal{R} \subseteq \mathcal{K}$, we have $K^* \subseteq R^*$. Now, let $B \in \mathcal{K}$. Then, by (40), $B \cap R^* \in \mathcal{R}$ and so, by the definition of R^*, $B \supseteq R^*$. Hence $K^* \supseteq R^*$ which gives the desired result.

(b) We first observe that, by A6 and part (a), relation (21) holds for $A = R^*$. Now, let A be any closed set as in part (b) of the theorem. Since the intersection of all sets belonging to $\mathcal{K}$ equals R^* as proved in part (a), it suffices to show that $A \in \mathcal{K}$. Taking $B = A_{z_1}$ in (21) yields $v(A.[z_1']z_1; x) \leqq v([z_1']z_1; x)$ for all $x \in A_{z_1}$. Next, by (32) and (31) with $V = A_{z_1}$, this inequality holds for all $x \in X$. This shows $A \in \mathcal{K}$ which ends the proof.

(c) Let the set A satisfy the assumptions of part (c). Then, by $R^* \subseteq A_{z_1}$, we have $v(A.[z_1']z_1; x) \leqq v(R^*.[z_1']z_1; x)$ for all $x \in A_{z_1}$. Further, by part (b), this inequality also holds with the inequality sign reversed. Hence $v(A.[z_1']z_1; x) =$

$v(R^*.[z_1']z_1; x)$ for all $x \in A_{z_1'}$. Next, using (32) with $V = A_{z_1'}$, we find that this equality holds for all $x \in X$. Since $R^* \in \mathscr{R}$, it now follows from A6(a) that $A \in \mathscr{R}$. We now obtain part (c) from Lemma 6.

References

[1] BREIMAN, L. (1968) *Probability*. Addison-Wesley, Reading, Massachusetts.

[2] DE LEVE, G. (1964) *Generalized Markovian Decision Processes, Part I: Model and Method. Part II: Probabilistic background.* Mathematical Centre Tracts No. 3 and 4, Mathematisch Centrum, Amsterdam.

[3] DE LEVE, G., TIJMS, H. C. AND WEEDA, P. J. (1970) *Generalized Markovian Decision Processes Applications.* Mathematical Centre Tract No. 5, Mathematisch Centrum, Amsterdam.

[4] DE LEVE, G., FEDERGRUEN, A. AND TIJMS, H. C. (1977) A general Markov decision method II: Applications. *Adv. Appl. Prob.* **9**,

[5] DE LEVE, G., FEDERGRUEN, A. AND TIJMS, H. C. (1977) *Generalized Markovian Decision Processes, Revisited.* Mathematical Centre Tract, Mathematisch Centrum, Amsterdam (to appear).

[6] DERMAN, C. AND VEINOTT, A. F., JR. (1967) A solution to a countable system of equations arising in Markovian decision processes. *Ann. Math. Statist.* **38**, 582–584.

[7] DERMAN, C. (1970) *Finite State Markovian Decision Processes.* Academic Press, New York.

[8] FELLER, W. (1957) *An Introduction to Probability Theory and its Applications.* Vol. 1, 2nd edn. Wiley, New York.

[9] FELLER, W. (1966) *An Introduction to Probability Theory and its Applications.* Vol. 2. Wiley, New York.

[10] HOWARD, R. A. (1960) *Dynamic Programming and Markov Processes.* M.I.T. Press, Cambridge, Mass.

[11] JAIN, N. C. (1966) Some limit theorems for general Markov processes. *Z. Wahrscheinlichkeitsth.* **5**, 206–223.

[12] JEWELL, W. S. (1963) Markov revewal programming, I: Formulation, finite return models. II: Infinite return models. *Opns. Res.* **6**, 938–972.

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

[14] Royden, H. L. (1968) *Real Analysis.* 2nd. edn. Macmillan, New York.

[15] WEEDA, P. J. (1974) Some computational experiments with a special generalized Markov programming model. Report BW 37/74, Mathematisch Centrum, Amsterdam.

[16] WEEDA, P. J. (1976) *Generalized Markov Programming and Markov Renewal Programming.* Mathematical Centre Tract, Mathematisch Centrum, Amsterdam. (To appear).