The Annals of Statistics
1976, Vol. 4, No. 6, 1219–1235

CONTINUOUS TIME CONTROL OF MARKOV PROCESSES ON AN ARBITRARY STATE SPACE: DISCOUNTED REWARDS

BY BHARAT T. DOSHI

Rutgers University

The paper deals with continuous time Markov decision processes on a fairly general state space. The rewards are continuously discounted at rate $\alpha > 0$. A set of conditions is shown to be necessary and sufficient for a policy to be optimal. For the special case of time independent reward function and under the assumption that the action space is finite a policy improvement algorithm is proposed and its convergence to an optimal policy is proved.

1. Introduction. We study continuous time Markov decision processes on a general state space. Such a decision process can be described as follows. $\{X_t; t \geqq 0\}$ is a stochastic process on a state space $\mathscr{X}$. At each of the specified time epochs $t \in \mathscr{T}$ the state X_t is observed and based on the history of the process up to time t, an action a_t is chosen from the action space $\mathscr{A}$. The actions $\{a_t; t \in \mathscr{T}\}$ interact with chance environments in determining the evolution of the process $\{X_t; t \geqq 0\}$. We assume that given the present state X_t and action a_t, the evolution of the process $\{X_s; s \geqq 0\}$ until the next decision is made is stochastically independent of the past. Because of this Markov property the resulting decision process is called a Markov decision process. At time t, if the state is x and action a is chosen, then the reward is obtained at the rate $r(t, x, a)$. A policy π is a measurable rule for choosing actions. That is, given the time t and the state x_t at time t, π prescribes action a_t to be chosen. For each policy π and initial state x we define its economic effectiveness by the long run total expected discounted return. A policy maximizing this total expected discounted return is called optimal. We seek to find the conditions under which an optimal policy exists. We also seek to characterize an optimal policy, if one exists.

Much of the earlier work in this area was done by Blackwell [1, 2] and Strauch [27]. They, however, restricted themselves to discrete time parameter case. That is, $\mathscr{T} = \{0, 1, 2, \cdots\}$. Hinderer [9] gives an extensive account of Markov decision processes with discrete time parameter.

Miller [17, 18] considered Markov decision processes with continuous time parameter. This is, $\mathscr{T} = [0, \infty)$. He restricted attention to finite state space case. Later, Kakumanu [13] extended his results to the case of countable state space. Markov decision processes with continuous time parameter and arbitrary

Received February 1975; revised March 1976.

AMS 1970 *subject classifications.* 49C15, 60K99.

Key words and phrases. Continuous time Markov decision processes, optimal discounted return function, optimal policy.

state space have not been studied in its generality. Special cases have been considered by various authors. Stone [26] and Pliska [21] dealt with controlled jump processes. Mandl [17], Kushner [15], Fleming [7] and Pliska [21] studied controlled diffusion processes. The methods used for the controlled diffusion processes rely on the properties of the second order differential operator and cannot be easily extended to more general processes.

In this paper we deal with continuous time Markov decision processes on a fairly general state space. No assumptions are made about the specific nature of the controlled process. Our approach is similar in spirit to that of Kakumanu [13]. The results obtained here will include as special cases those obtained by Kakumanu [13] and Pliska [21] for jump processes, because the infinitesimal operator A defined below reduces to the infinitesimal generator Q of [13] and [21]. With appropriate modifications due to boundary conditions they will also include the results obtained for the controlled diffusion processes.

In Section 3 we put the control problem described above in the formal framework of dynamic programming. A useful apparatus connecting a time-homogeneous Markov process with a contraction semigroup of bounded linear operators is developed in Section 2. A necessary and sufficient condition for a policy π^* to be optimal is derived in Section 4 which also shows the existence and uniqueness of a solution to the dynamic programming functional equation. The rest of this paper is devoted to an important special case of time independent reward function. In this case it is shown that the existence of an optimal policy implies the existence of a stationary optimal policy. For the problems with a finite action space an algorithm is presented to generate successively improving stationary policies. Under appropriate assumptions this algorithm is shown to converge to an optimal policy. This also provides a constructive proof of the existence of a stationary optimal policy when $\mathscr{A}$ is finite. When the action space $\mathscr{A}$ is countable the results for the finite action space case are used to show the existence of a solution to the dynamic programming functional equation and so that of an ε-optimal policy for any $\varepsilon > 0$.

2. Contraction semigroup associated with a time-homogeneous Markov process. The material in this section follows from Dynkin [5] where the reader is referred to for details.

Let Z be a complete separable metric space with the usual topology τ_Z and the Borel σ-algebra β_Z on it. For each $z \in Z$, let $(\Omega, \mathscr{F}, P_z)$ be a probability space. Let $\{Z_t; t \geqq 0\}$ be a time-homogeneous Markov process on Z with $(\Omega, \mathscr{F}, P_z)$ as the underlying probability space.

DEFINITION 2.1. The transition function P of the process $\{Z_t; t \geqq 0\}$ is defined by

$$P(z; t, \Gamma) = P_z\{Z_t \in \Gamma\}(z \in Z, \Gamma \in \beta_Z, t \geqq 0) .$$

Let $B(Z)$ denote the set of all bounded β_Z-measurable function on Z. With respect to the supnorm

$$||f|| = \sup_{z\in Z} |f(z)|$$

and the usual linear operations, $B(Z)$ is a Banach space.

DEFINITION 2.2. For each $t \geqq 0$, we define an operator T_t: $B(Z) \to B(Z)$ by

$$\begin{aligned} T_t f(z) &= E_z[f(Z_t)] = E[f(Z_t) \mid Z_0 = z] \\ &= \int_Z f(y) P(z; t, dy) \end{aligned}$$

for $f \in B(Z)$ and $z \in Z$.

The family $\{T_t; t \geqq 0\}$ is a contraction semigroup of bounded linear operators on $B(Z)$. Dynkin [5] gives some important properties of this semigroup [also see [4], Appendix I].

DEFINITION 2.3. Let $\{f_n; n \geqq 1\}$ be a sequence in $B(Z)$. We say that f_n converges to $f \in B(Z)$ in a *weak sense* if

(i) $\lim_{n\to\infty} f_n(z) = f(z)$ for each $z \in Z$, and
(ii) $\{||f_n||; n \geqq 1\}$ is a bounded sequence.

If f_n converges to f in a weak sense we write $w \lim_{n\to\infty} f_n = f$.

Let B_0 be a subset of $B(Z)$ such that for $f \in B_0$, $w \lim_{t\downarrow 0+} T_t f = f$.

DEFINITION 2.4. The *weak infinitesimal operator* A of a contraction semigroup $\{T_t; t \geqq 0\}$ is defined by

$$Af = w \lim_{t\downarrow 0+} \frac{T_t f - f}{t}$$

for all $f \in B_0$ such that the limit on the right-hand side exists and belongs to B_0. Let $\mathscr{D}(A) \subseteqq B_0$ denote the set of such functions.

3. **Formulation and basic assumptions.** We now formulate the control problem described in Section 1 as a dynamic programming problem. It is characterized by the following objects:

(a) The *state space* $\mathscr{X}$. This is assumed to be a nonempty Borel subset of a complete separable metric space. Let $\tau_{\mathscr{X}}$ be the usual topology on $\mathscr{X}$ and let $\beta_{\mathscr{X}}$ be the σ-algebra of Borel subsets of $\mathscr{X}$.

(b) The *action space* $\mathscr{A}$. This is also a nonempty Borel subset of a complete separable metric space. Let $\tau_{\mathscr{A}}$ and $\beta_{\mathscr{A}}$ denote the corresponding topology and σ-algebra, respectively.

(c) The *set* $\mathscr{T}$ *of decision epochs*. Since the decisions are taken continuously $\mathscr{T} = [0, \infty)$. Let τ_t and β_t be the corresponding topology and σ-algebra, respectively.

(d) The *law of motion*. Corresponding to each action $a \in \mathscr{A}$, there exists a weak infinitesimal operator A_a. If at any time t_0 action a is chosen, then the stochastic behaviour of $\{X_t; t \geqq 0\}$ at time t_0 is completely determined by x_{t_0} and A_a.

(e) The *reward rate function r*. This is assumed to be a measurable function on $[0, \infty) \times \mathscr{X} \times \mathscr{A}$.

(f) The set D_A of all *admissible policies*. A policy and its types are discussed below.

In this paper we restrict ourselves to (nonrandomized) *Markov* policies. A Markov policy is a $\beta_t \times \beta_{\mathscr{X}}$-measurable function on $[0, \infty) \times \mathscr{X}$ into $\mathscr{A}$ such that $\pi(t, x)$ is the action prescribed by π when state x is observed at time t. Let D_M denote the set of all Markov policies. A Markov policy is called stationary if it is independent of time. That is $\pi(t, x) = \pi(0, x) = \pi(x)$ for all $t \geqq 0$, $x \in \mathscr{X}$. Let D_S be the set of all stationary policies.

In a given application various reasons may force the controller to restrict the choice of his policy from a subset D_A of D_M. D_A is called the set of admissible policies. In this paper we shall assume that for each $\pi \in D_A$ there exists a version of stochastic process $\{X_t; t \geqq 0\}$ with the following properties:

(i) $\{X_t; t \geqq 0\}$ is a strongly measurable, strong Markov process.

(ii) The stochastic behaviour of the Markov process $\{X_t; t \geqq 0\}$ at any time t_0 is determined by the weak infinitesimal operator $A_{\pi(t_0, x_{t_0})}$.

(iii) Almost all sample paths of $\{X_t; t \geqq 0\}$ are right continuous with left limits and have only finitely many discontinuities in any finite interval of time.

Let P_π denote the transition probability function of $\{X_t; t \geqq 0\}$ under policy π. That is, $P_\pi(s, x; s + t, \Gamma) = P_\pi\{X_{t+s} \in \Gamma \mid X_s = x\}$ for all $s \geqq 0$, $t \geqq 0$, $x \in \mathscr{X}$ and $\Gamma \in \beta_{\mathscr{X}}$.

The Markov process $\{X_t; t \geqq 0\}$ induced by a policy $\pi \in D_A$ is not time-homogeneous except when π is a stationary policy. However, by properly expanding the state space we can get a time-homogeneous Markov process for any $\pi \in D_A$. More specifically, let (Z, β_Z) denote the product space $([0, \infty), \beta_t) \times (\mathscr{X}, \beta_{\mathscr{X}})$. For any policy $\pi \in D_A$, the bivariate process $\{(t, X_t); t \geqq 0\}$ is a time-homogeneous Markov process with transition function H_π given by

$$H_\pi(s, x; t, \tau, \Gamma) = P_\pi(s, x; s + t, \Gamma)\delta(\tau - s - t) \qquad s \geqq 0, t \geqq 0, \tau \geqq 0, \tag{3.1}$$

where $\delta(y) = 0$ if $y < 0$, and $= 1$ if $y \geqq 0$. When convenient, we shall use z to represent $(t, x) \in [0, \infty) \times \mathscr{X}$. In this case the process $\{(t, X_t); t \geqq 0\}$ will be represented by $\{Z_t; t \geqq 0\}$. Since $\pi \in D_A$ induces a time-homogeneous Markov process $\{Z_t; t \geqq 0\}$ we can associate with it, a contraction semigroup $\{T_t^\pi; t \geqq 0\}$ and its weak infinitesimal operator A_π. The sets B_0^π and $\mathscr{D}(A_\pi)$ have obvious meanings.

Unless otherwise stated, the continuous time Markov decision process under study is assumed to satisfy the following:

ASSUMPTION 1.

(a) There exists a nonempty subset B_0 of $B(Z)$ such that $B_0 \subset \bigcap_{\pi \in D_A} B_0^\pi$.

(b) There exists a nonempty subset $\mathscr{D}(A)$ of $B(Z)$ such that $\mathscr{D}(A) \subset \bigcap_{\pi \in D_A} \mathscr{D}(A_\pi)$.

(c) For each $\pi \in D_A$, the resulting Markov process $\{Z_t; t \geqq 0\}$ is conservative. That is, $T_t^\pi 1 = 1$ and $A_\pi 1 = 0$.

(d) There exists an $M < \infty$ such that

$$|r(t, x, a)| \leqq M$$

for all $t \geqq 0$, $x \in \mathscr{X}$ and $a \in \mathscr{A}$.

(e) For any $\pi \in D_A$, $r_\pi \in B_0$ where $r_\pi \colon Z \rightarrow R$ is defined by $r_\pi(t, x) = r(t, x, \pi(t, x))$ for $t \geqq 0$ and $x \in \mathscr{X}$.

Let $\alpha > 0$ be the discount rate. We discount the rewards continuously at a constant rate α. Suppose we start at time $t \geqq 0$ in state $x \in \mathscr{X}$ and use policy $\pi \in D_A$. The *total expected discounted return* is then given by

$$V_\pi(t, x) = E_\pi[\textstyle\int_t^\infty e^{-\alpha(\tau - t)} r(\tau, X_\tau, \pi(\tau, X_\tau))\, d\tau \mid X_t = x] . \tag{3.2}$$

By Fubini's theorem the interchange of expectation and integral is justified. This leads to

$$\begin{aligned} V_\pi(t, x) &= \textstyle\int_t^\infty e^{-\alpha(\tau - t)} E_\pi[r(\tau, X_\tau, \pi(\tau, X_\tau)) \mid X_t = x]\, d\tau \\ &= \textstyle\int_0^\infty e^{-\alpha\tau} T_\tau^\pi r_\pi(t, x)\, d\tau . \end{aligned} \tag{3.3}$$

DEFINITION 3.1. The optimal discounted return function $V_* \colon [0, \infty) \times \mathscr{X} \rightarrow R$ is defined by

$$V_*(t, x) = \sup_{\pi \in D_A} V_\pi(t, x) \qquad t \geqq 0,\ x \in \mathscr{X}. \tag{3.4}$$

DEFINITION 3.2. A policy $\pi^* \in D_A$ is said to be *α-optimal* or simply *optimal* in D_A if $V_{\pi^*}(t, x) = V_*(t, x)$ for all $t \geqq 0$ and $x \in \mathscr{X}$.

DEFINITION 3.3. For a given $\varepsilon > 0$, a policy $\pi^* \in D_A$ is said to be *ε-optimal* if $V_{\pi^*}(t, x) \geqq V_*(t, x) - \varepsilon$ for all $t \geqq 0$ and $x \in \mathscr{X}$.

4. Conditions for optimality. We begin by showing that the discounted return function V_π for any policy $\pi \in D_A$ is the unique solution in $\mathscr{D}(A_\pi)$ of a functional equation. This functional equation is of great importance because it depends on the controlled process $\{Z_t : t \geqq 0\}$ only through its infinitesimal operator A_π.

THEOREM 4.1. *For any policy $\pi \in D_A$, V_π is the unique solution in $\mathscr{D}(A_\pi)$ of the equation*

$$(\alpha I - A_\pi)V = r_\pi . \tag{4.1}$$

That is,

$$\alpha V = r_\pi + A_\pi V .$$

PROOF. The proof may be found in Dynkin [5, Theorem 1.7].

We now establish the necessity and sufficiency of a dynamic programming condition for a policy π^* to be optimal in D_A. We shall assume that the set D_A of admissible policies satisfies the following:

ASSUMPTION 2.

(a) Suppose $\pi \in D_A$. Let a policy π' be defined by

$$\pi'(t, x) = \pi(t_0, x) \qquad\qquad x \in \mathscr{X},\ t \geqq 0$$

for some $t_0 \geqq 0$. Then $\pi' \in D_A$.

(b) Suppose $\pi \in D_A$ and $\pi' \in D_A$. For some $t_0 \geqq 0$ and $\tau_0 > 0$ let π'' be defined by

$$\begin{aligned} \pi''(t, x) &= \pi(t, x) && \text{if} \quad 0 \leqq t < t_0\,, \quad t_0 + \tau_0 \leqq t < \infty \\ &= \pi'(t, x) && \text{if} \quad t_0 \leqq t < t_0 + \tau_0\,. \end{aligned}$$

Then $\pi'' \in D_A$. We say that the set D_A is *closed under* (*time*) *interval exchange* if it satisfies this assumption.

The following lemma was proved by Kakumanu [13] in the countable state space case. Let 1 denote the unit function on Z. That is

$$1(z) = 1 \tag{4.2}$$

for all $z \in Z$.

LEMMA 4.1. *Suppose there exists a function $V \in \mathscr{D}(A)$ and a nonnegative constant ε such that*

$$\alpha V \underset{(\leqq)}{\geqq} r_\pi + A_\pi V \underset{(+)}{-} \varepsilon 1\,. \tag{4.3}$$

Then

$$V \underset{(\leqq)}{\geqq} V_\pi \underset{(+)}{-} \varepsilon 1/\alpha\,. \tag{4.4}$$

PROOF. From (4.3) we have

$$(\alpha I - A_\pi)V \underset{(\leqq)}{\geqq} r_\pi \underset{(+)}{-} \varepsilon 1\,.$$

Since $(\alpha I - A_\pi)^{-1}$ is a monotone operator the above inequality may be written as

$$\begin{aligned} V &\underset{(\leqq)}{\geqq} (\alpha I - A_\pi)^{-1}(r_\pi - \varepsilon 1) \\ &= (\alpha I - A_\pi)^{-1}((\alpha I - A_\pi)V_\pi \underset{(+)}{-} \varepsilon 1) \\ &= V_\pi \underset{(+)}{-} \varepsilon 1/\alpha\,. \end{aligned}$$

This proves the lemma.

LEMMA 4.2. *If for some $\pi \in D_A$, $t \geqq 0$ and $x \in \mathscr{X}$, there exist $V \in \mathscr{D}(A_\pi)$, $f \in \mathscr{D}(A_\pi)$ and a $\delta > 0$ such that*

$$r_\pi(t, x) + A_\pi V(t, x) \gtrless f(t, x) \pm \delta\,, \tag{4.5}$$

then there exists a $\tau_0 > 0$ such that

$$T_\tau^\pi r_\pi(t, x) + T_\tau^\pi A_\pi V(t, x) \gtrless T_\tau^\pi f(t, x) \pm \delta \qquad 0 \leqq \tau < \tau_0\,. \tag{4.6}$$

PROOF. By Assumption 1(e) $r_\pi \in B_0 \subset B_0^\pi$ for all $\pi \in D_A$. By hypothesis of the lemma and by definition of $\mathscr{D}(A_\pi)$ we have $A_\pi V \in B_0^\pi$ and $f \in \mathscr{D}(A_\pi) \subset B_0^\pi$. The set B_0^π is closed under addition. Therefore,

$$r_\pi A_\pi V - f \pm \delta 1 \in B_0^\pi .$$

The lemma now follows from the definition of B_0^π.

THEOREM 4.2. *Suppose $\pi^* \in D_A$ and $V_{\pi^*} \in \mathscr{D}(A)$. Then a necessary and sufficient condition for π^* to be optimal in D_A is that*

$$\alpha V_{\pi^*}(t, x) = \sup_{\pi \in D_A} \{r_\pi(t, x) + A_\pi V_{\pi^*}(t, x)\} \qquad t \geqq 0,\ x \in \mathscr{X}. \tag{4.7}$$

PROOF. Suppose π^* satisfies (4.7). It follows that

$$\alpha V_{\pi^*} \geqq r_\pi + A_\pi V_{\pi^*} \qquad \pi \in D_A .$$

Lemma 4.1 now implies that

$$V_{\pi^*} \geqq V_\pi \qquad \pi \in D_A ,$$

which proves the sufficiency.

To establish the necessity suppose there exists a $\pi^* \in D_A$ such that $V_{\pi^*} \in \mathscr{D}(A)$ and

$$V_{\pi^*} = V_* = \sup_{\pi \in D_A} V_\pi .$$

From Theorem 4.1 it follows that

$$\alpha V_{\pi^*}(t, x) \leqq \sup_{\pi \in D_A} \{r_\pi(t, x) + A_\pi V_{\pi^*}(t, x)\} \qquad t \geqq 0,\ x \in \mathscr{X}. \tag{4.8}$$

Suppose there exist $t_0 \geqq 0$, $x_0 \in \mathscr{X}$ and $\pi \in D_A$ such that

$$\alpha V_{\pi^*}(t_0, x_0) < r_\pi(t_0, x_0) + A_\pi V_{\pi^*}(t_0, x_0) .$$

Then there exists a $\delta > 0$ such that

$$\alpha V_{\pi^*}(t_0, x_0) < r_\pi(t_0, x_0) + A_\pi V_{\pi^*}(t_0, x_0) - \delta . \tag{4.9}$$

From Lemma 4.2 it follows that for some $\tau_0 > 0$

$$T_\tau^\pi \alpha V_{\pi^*}(t_0, x_0 < T_\tau^\pi r_\pi(t_0, x_0) + T_\tau^\pi A_\pi V_{\pi^*}(t_0, x_0) - \delta \qquad 0 \leqq \tau < \tau_0 . \tag{4.10}$$

Let a policy π' be defined by

$$\begin{aligned} \pi'(t, x) &= \pi(t, x) && \text{if } \ t_0 \leqq t < t_0 + \tau_0 \\ &= \pi^*(t, x) && \text{if } \ 0 \leqq t < t_0 , \quad t_0 + \tau_0 \leqq t < \infty . \end{aligned}$$

By Assumption 2(b) $\pi' \in D_A$. From (4.8) and (4.10) we obtain

$$-\frac{d^+}{d\tau}(e^{-\alpha\tau} T_\tau^{\pi'} V_{\pi^*}(t_0, x_0)) < e^{-\alpha\tau} T_\tau^{\pi'} r_{\pi'}(t_0, x_0) - e^{-\alpha\tau}\delta \qquad \text{if } \ 0 \leqq \tau < \tau_0$$

and

$$-\frac{d^+}{d\tau}(e^{-\alpha\tau} T_\tau^{\pi'} V_{\pi^*}(t_0, x_0)) \leqq e^{-\alpha\tau} T_\tau^{\pi'} r_{\pi'}(t_0, x_0) \qquad \text{if } \ \tau \geqq \tau_0 .$$

Integrating from $\tau = 0$ to ∞ the above inequalities reduce to

$$\begin{aligned} V_{\pi^*}(t_0, x_0) &\leqq V_{\pi'}(t_0, x_0) - (\delta/\alpha)(1 - e^{-\alpha\tau}0) \\ &< V_{\pi'}(t_0, x_0) . \end{aligned}$$

This contradicts the fact that π^* is optimal in D_A. The necessity is thus proved by contradiction.

The above theorem is useful only when an optimal policy exists. There are situations where an optimal policy does not exist. Some Markov decision processes with countably or uncountably infinite action space are of this type. The optimal discounted return function V_* is, however, well defined by equation (3.4) in these situations. We now study this optimal discounted return function and also the question of an ε-optimal policy. Some additional assumptions are needed to obtain useful results.

ASSUMPTION 3. If $V \in \mathscr{D}(A)$ and $f \in \mathscr{D}(A)$ satisfy

$$(4.11) \qquad f(t, x) = \sup_{\pi \in D_A} \{r_\pi(t, x) + A_\pi V(t, x)\} \qquad t \geqq 0,\ x \in \mathscr{X},$$

then for any given $\varepsilon > 0$, there exists a $\pi' \in D_A$ such that

$$(4.12) \qquad r_{\pi'}(t, x) + A_{\pi'} V(t, x) + \varepsilon > f(t, x) \qquad t \geqq 0,\ x \in \mathscr{X}.$$

ASSUMPTION 4. Let μ be any probability measure on $\mathscr{X}$. Then for given $\varepsilon > 0$ and $t \geqq 0$, there exists a policy $\pi_{t,\varepsilon} \in D_A$ such that

$$(4.13) \qquad \mu\{x : V_{\pi_{t,\varepsilon}}(t, x) \geqq V_*(t, x) - \varepsilon\} = 1 .$$

ASSUMPTION 5. Let D' be any subset of D_A. Let $f \in \mathscr{D}(A)$ and $V \in \mathscr{D}(A)$. If for some $t_0 \geqq 0$ and $x_0 \in \mathscr{X}$, there exists a $\delta > 0$ such that

$$(4.14) \qquad r_\pi(t_0, x_0) + A_\pi V(t_0, x_0) \pm \delta \lessgtr f(t_0, x_0) \qquad \pi \in D' ,$$

then there exists a $\tau_0 > 0$ and a $\delta_1 > 0$ satisfying

$$(4.15) \qquad T_\tau^\pi r_\pi(t_0, x_0) + T_\tau^\pi A_\pi V(t_0, x_0) \pm \delta_1 \lessgtr T_\tau^\pi f(t_0, x_0) \qquad 0 \leqq \tau < \tau_0,\ \pi \in D' .$$

REMARK. Maitra [16] has investigated conditions under which Assumption 3 holds. For discrete time parameter problem Blackwell [2] and Strauch [27] have shown that Assumption 4 holds under fairly general conditions. Schal [25] has recently proved the same result using an entirely different approach. Schal's approach appears to be more promising as far as the extension to continuous time parameter case is concerned. Assumption 5 imposes a uniformity property on the result of Lemma 4.3. It is clearly satisfied when $\mathscr{A}$ is finite.

THEOREM 4.3. *Suppose $V_* = \sup_{\pi \in D_A} V_\pi \in \mathscr{D}(A)$. Then V_* is the unique solution in $\mathscr{D}(A)$ of*

$$(4.16) \qquad \alpha V(t, x) = \sup_{\pi \in D_A} \{r_\pi(t, x) + A_\pi V(t, x)\} \qquad t \geqq 0,\ x \in \mathscr{X},$$

and for any $\varepsilon > 0$, there exists an ε-optimal policy in D_A.

PROOF. We first prove that (4.16) has a unique solution, if any, in $\mathscr{D}(A)$.

Suppose $V \in \mathscr{D}(A)$ satisfies (4.16). Then by Lemma 4.1

$$V \geqq V_* . \tag{4.17}$$

Let $\varepsilon > 0$ be given. By Assumption 3 there exists a policy $\pi_\varepsilon \in D_A$ satisfying

$$\alpha V < r_{\pi_\varepsilon} + A_{\pi_\varepsilon} V + \varepsilon\alpha 1 .$$

Lemma 4.1 now implies that

$$\begin{aligned} V &\leqq V_{\pi_\varepsilon} + \varepsilon 1 \\ &\leqq V_* + \varepsilon 1 \end{aligned} \tag{4.18}$$

since $\varepsilon > 0$ can be made arbitrarily small, we conclude that

$$V \leqq V_* . \tag{4.19}$$

Combining (4.17) and (4.19) we obtain

$$V = V_* ,$$

and from (4.18) it follows that π_ε is an ε-optimal policy.

We now show that V_* satisfies equation (4.16). Suppose for some $t_0 \geqq 0$ and $x_0 \in \mathscr{X}$,

$$\alpha V_*(t_0, x_0) < \sup_{\pi \in D_A} \{r_\pi(t_0, x_0) + A_\pi V_*(t_0, x_0)\} . \tag{4.20}$$

That is, there exist a policy $\pi \in D_A$ and a $\delta > 0$ satisfying

$$\alpha V_*(t_0, x_0) < r_\pi(t_0, x_0) + A_\pi V_*(t_0, x_0) - \delta .$$

By Lemma 4.2 there exists a $\tau_0 > 0$ such that

$$\alpha e^{-\alpha\tau} T_\tau{}^\pi V_*(t_0, x_0) < e^{-\alpha\tau} T_\tau{}^\pi r_\pi(t_0, x_0) + e^{-\alpha\tau} T_\tau{}^\pi A_\pi V_*(t_0, x_0) - \delta e^{-\alpha\tau} \tag{4.21}$$

for $0 \leqq \tau < \tau_0$. Let $\varepsilon > 0$. By Assumption 4 there exists a policy $\pi_{t_0+\tau_0,\varepsilon}$ with

$$e^{-\alpha\tau_0} T_{\tau_0}^\pi V_*(t_0, x_0) \leqq e^{-\alpha\tau_0} T_{\tau_0}^\pi V_{\pi_{t_0+\tau_0,\varepsilon}}(t_0, x_0) + e^{-\alpha\tau_0}\varepsilon . \tag{4.22}$$

We define a policy π^* by

$$\begin{aligned} \pi^*(t, x) &= \pi(t, x) && \text{if} \quad 0 \leqq t < t_0 + \tau_0 \\ &= \pi_{t_0+\tau_0,\varepsilon}(t, x) && \text{if} \quad t_0 + \tau_0 \leqq t < \infty . \end{aligned}$$

By Assumption 2(b) $\pi^* \in D_A$. Also from (4.21) and (4.22) we have

$$V_*(t_0, x_0) \leqq V_{\pi^*}(t_0, x_0) - (\delta/\alpha)(1 - e^{-\alpha\tau_0}) + \varepsilon e^{-\alpha\tau_0} .$$

Choosing $\varepsilon < (\delta/\alpha)(1 - e^{-\alpha\tau_0})/e^{-\alpha\tau_0}$ we obtain

$$V_*(t_0, x_0) < V_{\pi^*}(t_0, x_0) .$$

This contradicts the fact that V_* is the optimal return function. So

$$\alpha V_*(t, x) \geqq \sup_{\pi \in D_A} \{r_\pi(t, x) + A_\pi V_*(t, x)\} \qquad t \geqq 0,\ x \in \mathscr{X} . \tag{4.23}$$

Suppose there exist $t_0 \geqq 0$ and $x_0 \in \mathscr{X}$ such that

$$\alpha V_*(t_0, x_0) > \sup_{\pi \in D_A} \{r_\pi(t_0, x_0) + A_\pi V_*(t_0, x_0)\} . \tag{4.24}$$

That is, there exists a $\delta > 0$ satisfying

$$\alpha V_*(t_0, x_0) > r_\pi(t_0, x_0) + A_\pi V_*(t_0, x_0) + \delta \qquad \pi \in D_A .$$

Assumption 5 implies the existence of a $\delta_1 > 0$ and a $\tau_0 > 0$ with

$$\alpha e^{-\alpha\tau} T_\tau^{\ \pi} V_*(t_0, x_0) > e^{-\alpha\tau} T_\tau^{\ \pi} r_\pi(t_0, x_0) + e^{-\alpha\tau} T_\tau^{\ \pi} V_*(t_0, x_0) + \delta_1 e^{-\alpha\tau} \tag{4.25}$$

for all $\pi \in D_A$ and $0 \leqq \tau < \tau_0$. From (4.23) and (4.25) we obtain

$$V_*(t_0, x_0) \geqq (\delta_1/\alpha)(1 - e^{\alpha\tau_0}) + V_\pi(t_0, x_0) \qquad \pi \in D_A .$$

This implies that

$$V_*(t_0, x_0) > V_*(t_0, x_0)$$

which is a contradiction. So

$$\alpha V_*(t, x) = \sup_{\pi \in D_A} \{r_\pi(t, x) + A_\pi V_*(t, x)\} \qquad t \geqq 0, x \in \mathscr{X} .$$

5. Time independent reward function. In this section we study the Markov decision processes with time-independent reward function. That is, the case in which $r(t, x, a)$ does not depend on. t. In fact, most earlier literature on Markov decision theory deals only with this special case. We first investigate the conditions which are sufficient for the existence of an optimal policy which is stationary.

Sufficient conditions.

THEOREM 5.1. *If r is independent of time, then the following hold*:

(a)

$$V_*(t, x) = V_*(0, x) = V_*(x) \qquad t \geqq 0, x \in \mathscr{X} . \tag{5.1}$$

(b) *If there exists a policy π' which is optimal in D_A with $V_* \in D_A$, then there exists a policy $\pi^* \in D_A \cap D_S$ which is optimal in D_A.*

PROOF. The proof is trivial. The details may be found in [4], Lemma 2.5 and Theorem 2.7.

In Theorem 4.3 we derived a functional equation the solution of which is the optimal return function. We now use this functional equation to characterize an optimal policy. We need an assumption similar to Assumption 3.

ASSUMPTION 6. $V\colon \mathscr{X} \to R$ belong to $\mathscr{D}(A)$. If for each $x \in \mathscr{X}$, there exists an action $a_x \in \mathscr{A}$ such that

$$r(x, a_x) + A_{a_x} V(x) = \sup_{a \in \mathscr{A}} \{r(x, a) + A_a V(x)\} , \tag{5.2}$$

then there exists a policy $\pi^* \in D_A \cap D_S$ satisfying

$$\begin{aligned} r_{\pi^*}(x) + A_{\pi^*} V(x) &= \sup_{\pi \in D_S} \{r_\pi(x) + A_\pi V(x)\} \\ &= \sup_{a \in \mathscr{A}} \{r(x, a) + A_a V(x)\} \qquad x \in \mathscr{X} . \end{aligned} \tag{5.3}$$

In order to avoid repetitions we shall assume that our selection of a_x is such that

$$\pi^*(x) = a_x \qquad x \in \mathscr{X} ,$$

where $\pi^* \in D_A \cap D_S$.

Under Assumption 6 the following is obvious: Suppose there exists a function $V \in \mathscr{D}(A)$ satisfying

$$(5.4) \qquad \alpha V(x) = \sup_{a \in \mathscr{A}} \{r(x, a) + A_a V(x)\} = \sup_{\pi \in D_A} \{r_\pi(t, x) + A_\pi V(x)\} \qquad t \geqq 0,\ x \in \mathscr{X}.$$

(a) If $\mathscr{A}$ is finite, then there exists a policy $\pi^* \in D_A \cap D_S$ which is optimal in D_A.

(b) If $r(x, a) + A_a V(x)$ is an upper semicontinuous function of a with respect to the topology $\tau_{\mathscr{A}}$ and if $\mathscr{A}$ is compact with respect to this topology, then there exists a policy $\pi^* \in D_A \cap D_S$ which is optimal in D_A.

An important requirement in the above is the existence of a solution to the functional equation (5.4). The existence can be established using Theorem 4.3. But the Assumption 4 used in the proof of Theorem 4.3 is difficult to be verified and so an alternative approach is desired. A computational method provides such an alternative by giving a constructive proof of the existence of a solution to (5.4). This is a version of the *policy improvement algorithm* originally suggested by Howard [10]. We now describe this algorithm and prove that it generates an improving sequence of stationary policies. We also seek to establish the conditions under which this iterative procedure converges to a stationary policy which is optimal in D_A. We shall assume that the action space $\mathscr{A}$ is finite.

ALGORITHM. Given a policy $\pi^1 \in D_A \cap D_S$ we generate a sequence $\{\pi^n; n \geqq 1\}$ of policies in $D_A \cap D_S$ by the policy improvement algorithm. An iteration of this algorithm is described below.

(a) After finding $\pi^n \in D_A \cap D_S$ we obtain the expected discounted return function for π^n from the value determination equations

$$(5.5) \qquad \alpha V_{\pi^n}(x) = r_{\pi^n}(x) + A_{\pi^n} V_{\pi^n}(x) \qquad x \in \mathscr{X}.$$

If $n \geqq 2$ and $V_{\pi^n}(x) = V_{\pi^{n-1}}(x)$ for all $x \in \mathscr{X}$, then we terminate the algorithm and conclude that π^{n-1} and π^n are optimal in D_A. Otherwise proceed to (b).

(b) A policy π^{n+1} is defined by

$$\pi^{n+1}(x) = a_x \qquad x \in \mathscr{X}$$

where for each $x \in \mathscr{X}$,

$$(5.6) \qquad r(x, a_x) + A_{a_x} V_{\pi^n}(x) = \sup_{a \in \mathscr{A}} \{r(x, a) + A_a V_{\pi^n}(x)\}.$$

Assumption 6 guarantees that we can choose $\{a_x; x \in \mathscr{X}\}$ such that $\pi^{n+1} \in D_A \cap D_S$. We now go back to step (a).

In the following theorem we establish that the policy improvement algorithm described above generates a successively improving sequence of stationary policies.

THEOREM 5.2. *Let $\pi^1 \in D_A \cap D_S$ be given and $\{\pi^n; n \geqq 1\}$ be a sequence of policies in $D_A \cap D_S$ generated by the policy improvement algorithm starting with π^1.*

(a) *If for some* $n \geqq 1$

$$r_{\pi^{n+1}}(x) + A_{\pi^{n+1}}V_{\pi^n}(x) = r_{\pi^n}(x) + A_{\pi^n}V_{\pi^n}(x) \qquad x \in \mathscr{X}, \tag{5.7}$$

then

$$V_{\pi^{n+1}}(x) = V_{\pi^n}(x) \qquad x \in \mathscr{X}.$$

Also, π^n *and* π^{n+1} *are optimal in* D_A.

(b) *For all* $n \geqq 1$

$$V_{\pi^{n+1}}(x) \geqq V_{\pi^n}(x) \qquad x \in \mathscr{X}.$$

If for some $x_0 \in \mathscr{X}$

$$r_{\pi^{n+1}}(x_0) + A_{\pi^{n+1}}V_{\pi^n}(x_0) > r_{\pi^n}(x_0) + A_{\pi^n}V_{\pi^n}(x_0)\,, \tag{5.8}$$

then

$$V_{\pi^{n+1}}(x_0) > V_{\pi^n}(x_0)\,.$$

PROOF.

(a) By (5.7), Theorem 4.1 and the algorithm

$$\begin{aligned}\alpha V_{\pi^{n+1}}(x) = \alpha V_{\pi^n}(x) &= \sup_{a\in\mathscr{A}} \{r(x, a) + A_a V_{\pi^n}(x)\} \\ &= \sup_{\pi\in D_A} \{r_\pi(t, x) + A_\pi V_{\pi^n}(x)\} \qquad x \in \mathscr{X}.\end{aligned}$$

So

$$V_{\pi^{n+1}} = V_{\pi^n} = V_*$$

by Theorem 4.2.

(b) By definition of π^{n+1} we have

$$\begin{aligned} r_{\pi^{n+1}}(x) + A_{\pi^{n+1}}V_{\pi^n}(x) &\geqq r_{\pi^n}(x) + A_{\pi^n}V_{\pi^n}(x) \\ &= \alpha V_{\pi^n}(x) \qquad x \in \mathscr{X}.\end{aligned} \tag{5.9}$$

Lemma 4.1 now implies that

$$V_{\pi^{n+1}}(x) \geqq V_{\pi^n}(x) \qquad x \in \mathscr{X}.$$

Next suppose (5.8) holds for some $x_0 \in \mathscr{X}$. Then there exists a $\delta > 0$ such that

$$\alpha V_{\pi^n}(x_0) < r_{\pi^{n+1}}(x_0) + A_{\pi^{n+1}}V_{\pi^n}(x_0) - \delta\,.$$

By Lemma 4.2 there exists a $\tau_0 > 0$ satisfying

$$\alpha e^{-\alpha\tau}T_\tau^{\pi^{n+1}}V_{\pi^n}(x_0) < e^{-\alpha\tau}T_\tau^{\pi^{n+1}}r_{\pi^{n+1}}(x_0) + e^{-\alpha\tau}T_\tau^{\pi^{n+1}}A_{\pi^{n+1}}V_{\pi^n}(x_0) - \delta e^{-\alpha\tau} \quad \text{if} \quad 0 \leqq \tau < \tau_0 \tag{5.10}$$

and

$$\alpha e^{-\alpha\tau}T_\tau^{\pi^{n+1}}V_{\pi^n}(x_0) \leqq e^{-\alpha\tau}T_\tau^{\pi^{n+1}}r_{\pi^{n+1}}(x_0) + e^{-\alpha\tau}T_\tau^{\pi^{n+1}}A_{\pi^{n+1}}V_{\pi^n}(x_0) \quad \text{if} \quad \tau \geqq \tau_0\,. \tag{5.11}$$

As in proof of Theorem 4.2, (5.10) and (5.11) lead to

$$\begin{aligned} V_{\pi^n}(x_0) &\leqq V_{\pi^{n+1}}(x_0) - (\delta/\alpha)(1 - e^{-\alpha\tau_0}) \\ &< V_{\pi^{n+1}}(x_0)\,.\end{aligned}$$

This proves the theorem.

Thus, starting with a policy $\pi^1 \in D_A \cap D_S$ the policy improvement algorithm generates a sequence $\{\pi^n;\, n \geqq 1\}$ in $D_A \cap D_S$ such that

$$V_{\pi^{n+1}}(x) \geqq V_{\pi^n}(x) \qquad n \geqq 1,\ x \in \mathscr{X}, \tag{i}$$

and

$$\text{(ii)} \qquad ||V_{\pi^n}|| \leqq M/\alpha \qquad n \geqq 1 .$$

Therefore, $w \lim_{n\to\infty} V_{\pi^n}$ exists and is bounded. Let $V\colon \mathscr{X} \to R$ be defined by

$$V = w \lim_{n\to\infty} V_{\pi^n} . \tag{5.12}$$

We now study the conditions under which V equals the optimal return function V_* and satisfies the functional equation (5.4).

ASSUMPTION 7. Since $\mathscr{A}$ is finite, Tykonoff's theorem [28] implies that the set $\mathscr{A}^{\mathscr{X}}$ is compact with respect to the *weak topology*. The set of all stationary policies is isomorphic to $\mathscr{A}^{\mathscr{X}}$ and so is compact with respect to the *topology of weak convergence*. We shall assume that the set $D_A \cap D_S$ is also compact with respect to this topology

ASSUMPTION 8.

(a) $\{V_\pi;\ \pi \in D_A \cap D_S\}$ is a uniformly bounded equicontinuous family of functions.

(b) For any $\pi^* \in D_A \cap D_S$, the family $\{A_{\pi^*}V_\pi;\ \pi \in D_A \cap D_S\}$ is uniformly bounded and equicontinuous.

(c) For each $x \in \mathscr{X}$, there exist (i) a compact subset E_x of $\mathscr{X}$ containing x, (ii) a real number $M_0 < \infty$, and (iii) a real number $h_0 > 0$ such that for any $\pi \in D_A$ and $t \geqq 0$

$$P_\pi(t, x;\ t + h, E_x) \geqq 1 - M_0 h \qquad 0 \leqq h < h_0 . \tag{5.13}$$

Under these assumptions we prove the following two lemmas which will be useful in establishing the convergence properties of the policy improvement algorithm.

LEMMA 5.1. *Suppose that* $V_\pi \in \mathscr{D}(A)$ *for all* $\pi \in D_A \cap D_S$. *If for some sequence* $\{\pi^n;\ n \geqq 1\}$ *in* $D_A \cap D_S$, $V = w \lim_{n\to\infty} V_{\pi^n}$, *then* $V \in \bigcap_{\pi \in D_A} B_0^\pi$.

PROOF. By Assumption 8(a) the family $\{V_{\pi^n};\ n \geqq 1\}$ is uniformly bounded and equicontinuous. It follows from Ascoli-Arzela theorem that there exists a subsequence $\{\pi^{n'}\}$ of $\{\pi^n;\ n \geqq 1\}$ such that $V_{\pi^{n'}} \to V$ uniformly on every compact subset of $\mathscr{X}$. For $\pi \in D_A$, $t \geqq 0$, $x \in \mathscr{X}$ and $h > 0$, we have

$$\begin{aligned} |T_h^\pi V(t, x) - V(x)| &\leqq |T_h^\pi V(t, x) - T_h^\pi V_{\pi^n}(t, x)| \\ &\quad + |T_h^\pi V_{\pi^n}(t, x) - V_{\pi^n}(x)| + |V_{\pi^n}(x) - V(x)| . \end{aligned} \tag{5.14}$$

Let M_0, h_0 and E_x be as defined in Assumption 8(c). Then (5.14) reduces to

$$\begin{aligned} &|T_h^\pi V(t, x) - V(x)| \\ &\quad \leqq |\textstyle\int_{E_x} P_\pi(t, x;\ t + h, dy)[V(y) - V_{\pi^n}(y)]| + \dfrac{2M_0hM}{\alpha} \\ &\quad + |T_h^\pi V_{\pi^n}(t, x) - V_{\pi^n}(x)| + |V_{\pi^n}(x) - V(x)| \\ &\qquad\qquad 0 \leqq h < h_0,\ n \geqq 1 . \end{aligned} \tag{5.15}$$

Suppose $\varepsilon > 0$ is given. Since the convergence of $V_{\pi^{n'}}$ to V is uniform on E_x and since $V_{\pi^{n'}} \in \mathscr{D}(A) \subset \bigcap_{\pi \in D_A} B_0^\pi$, there exists a $\delta_0 > 0$ with the following properties.

(i) There exists a positive integer n_0 such that

$$|V(y) - V_{\pi^{n'}}(y)| < \varepsilon/4 \qquad y \in E_x n' \geqq n_0\,,$$

(ii) $$|T_h^\pi V_{\pi^{n_0}}(t, x) - V_{\pi^{n_0}}(x)| < \varepsilon/4 \qquad 0 \leqq h < \delta_0\,,$$

(iii) $$\delta_0 < h_0 \quad \text{and} \quad \delta_0 < (\varepsilon\alpha)/(8M_0 M)\,.$$

Substituting (i), (ii) and (iii) into (5.15) we obtain

$$\begin{aligned} |T_h^\pi V(t, x) - V(x)| &\leqq \varepsilon/4 + (2M_0 M\varepsilon\alpha)/(8M_0 M\alpha) + \varepsilon/4 + \varepsilon/4 \\ &= \varepsilon \qquad\qquad 0 \leqq h < \delta_0\,. \end{aligned}$$

Thus

$$\lim_{h \downarrow 0+} T_h^\pi V(t, x) = V(x) \qquad x \in \mathscr{X},\ t \geqq 0\,.$$

From boundedness of V it follows that

$$w \lim_{h \downarrow 0+} T_h^\pi V = V\,.$$

So $V \in B_0^\pi$. Since π was arbitrary, $V \in \bigcap_{\pi \in D_A} B_0^\pi$.

LEMMA 5.2. *Under the hypothesis of Lemma* 5.1 *the following hold*:

(a) $V \in \bigcap_{D_A \cap D_S} \mathscr{D}(A_\pi)$.

(b) *If* $\pi^* \in D_A \cap D_S$, *then there exists a subsequence* $\{\pi^{n'}\}$ *of* $\{\pi^n,\ n \geqq 1\}$ *such that*

$$w \lim_{n' \to \infty} A_{\pi^*} V_{\pi^{n'}} = A_{\pi^*} V\,.$$

PROOF. The lemma follows from Assumption 8 (b) and the fact that the operator A_π is closed (see Dynkin [5, page 40]).

Using the above two lemmas we now prove the convergence of the policy improvement algorithm and the existence of a stationary policy which is optimal in D_A.

THEOREM 5.3. *Suppose that* $V_\pi \in \mathscr{D}(A)$ *for* $\pi \in D_A \cap D_S$, $\mathscr{A}$ *is finite and* $\{\pi^n;\ n \geqq 1\}$ *is a sequence in* $D_A \cap D_S$ *generated by the policy improvement algorithm. Then the following hold*:

(a) $V \in \mathscr{D}(A)$.

(b) $\alpha V(x) = \sup_{a \in \mathscr{A}} \{r(x, a) + A_a V(x)\} = \sup_{\pi \in D_A} \{r_\pi(t, x) + A_\pi V(x)\}$ $(t \geqq 0,\ x \in \mathscr{X})$.

(c) $V(x) = V_*(x)$ $(x \in \mathscr{X})$.

(d) *There exists a policy* $\pi^* \in D_A \cap D_S$ *which is optimal in* D_A.

(e) *There exists a subsequence* $\{\pi^{n'}\}$ *of* $\{\pi^n;\ n \geqq 1\}$ *such that* $\pi^{n'}$ *converges to* π^* *pointwise*.

PROOF. From Assumption 7 and the uniform boundedness of r and V_{π^n} it follows that there exists a subsequence $\{\pi^{n'}\}$ of $\{\pi^n;\ n \geqq 1\}$ and a policy $\pi^* \in D_A \cap D_S$ with the following properties:

(i) $\pi^{n'} \to \pi^*$ pointwise. That is, for each $x \in \mathscr{X}$ there exists an integer n_x such that $\pi^{n'}(x) = \pi^*(x)$ for $n' \geqq n_x$.

(ii) $r_{\pi^{n'}}(x) = r_{\pi^*}(x)$ for $x \in \mathscr{X}$, $n' \geqq n_x$.

(iii) $A_{\pi^{n'}}V(x) = A_{\pi^*}V(x)$ for $x \in \mathscr{X}$, $n' \geqq n_x$ and $V \in \bigcap_{\pi \in D_A \cap D_S} \mathscr{D}(A_\pi)$.

Therefore

$$w \lim_{n' \to \infty} \alpha V_{\pi^{n'}} = \alpha V \in \bigcap_{\pi \in D_A} B_0^\pi \subset B_0^{\pi^*}, \tag{5.16}$$

and

$$w \lim_{n' \to \infty} r_{\pi^{n'}} = r_{\pi^*} \in B_0^{\pi^*}. \tag{5.17}$$

By Theorem 4.1 and equations (5.16) and (5.17) we have

$$w \lim_{n' \to \infty} A_{\pi^{n'}} V_{\pi^{n'}} = \alpha V - r_{\pi^*} \in B_0^{\pi^*}.$$

It follows from simple modifications in the proof of Lemma 5.2 that $V \in \mathscr{D}(A_{\pi^*})$, and

$$\alpha V = r_{\pi^*} + A_{\pi^*} V. \tag{5.18}$$

From Theorem 4.1 and the hypothesis we now have

$$V = V_{\pi^*} \in \mathscr{D}(A).$$

This proves (a) and (e).

From (5.18) we also obtain

$$\alpha V(x) \leqq \sup_{a \in \mathscr{A}} \{r(x, a) + A_a V(x)\} \qquad x \in \mathscr{X}.$$

For each n'

$$\begin{aligned} \alpha V_{\pi^{n'}}(x) \geqq r(x, a) &+ A_a V_{\pi^{n'-1}}(x) + A_{\pi^{n'}} V_{\pi^{n'}}(x) \\ &- A_{\pi^{n'}} V_{\pi^{n'-1}}(x) \qquad x \in \mathscr{X},\ a \in \mathscr{A}. \end{aligned} \tag{5.19}$$

By arguments similar to those used in the proof of Lemma 5.2, the above inequality may be reduced to

$$\alpha V(x) \geqq r(x, a) + A_a V(x) \qquad x \in \mathscr{X},\ a \in \mathscr{A}.$$

That is,

$$\alpha V(x) \geqq \sup_{a \in \mathscr{A}} \{r(x, a) + A_a V(x)\} \qquad x \in \mathscr{X}. \tag{5.20}$$

From (5.18) and (5.20) we obtain

$$\alpha V(x) = \sup_{a \in \mathscr{A}} \{r(x, a) + A_a V(x)\} = r_{\pi^*}(x) + A_{\pi^*} V(x) \qquad x \in \mathscr{X}.$$

(b), (c) and (d) now follow from Theorem 4.2.

Countable action space. When the action space is not finite the policy improvement algorithm is not feasible in general. Therefore the existence of a solution to the functional equation (5.4) and of a stationary policy that is optimal in D_A cannot be established directly using this algorithm. However, when $\mathscr{A}$ is countable we can prove the existence of a solution to the functional equation (5.4) and that of a stationary ε-optimal policy. The approach is via the finite action space case. Suppose that the countable action space $\mathscr{A}$ is given by $\{1, 2, \cdots\}$.

For each $n \geqq 1$, let $\mathscr{A}_n$ be the subset $\{1, 2, \cdots, n\}$ of $\mathscr{A}$, and $D_S{}^n$ the set of stationary policies corresponding to the action space $\mathscr{A}_n$. We have already proved that V_n is the unique solution in $\mathscr{D}(A)$ of the equation (5.4). We now use this fact to establish the existence of a solution to (5.4) when $\mathscr{A}$ is countable.

THEOREM 5.4. *Suppose $V_\pi \in \mathscr{D}(A)$ for all $\pi \in D_A \cap D_S$, and $D_A \cap D_S{}^n$ is compact for each $n \geqq 1$. Then we have the following:*

(a) *There exists a function $V \in \bigcap_{\pi \in D_A \cap D_S} \mathscr{D}(A_\pi)$ satisfying* (5.4).

(b) $V(x) = \sup_{\pi \in D_A \cap D_S} V_\pi(x)$ $(x \in \mathscr{X})$.

(c) *For $\varepsilon > 0$, there exists a policy $\pi^* \in D_A \cap D_S$ which is ε-optimal in $D_A \cap D_S$.*

If $V \in \mathscr{D}(A)$, then for any $\varepsilon > 0$ there exists a policy $\pi^* \in D_A \cap D_S$ which is ε-optimal in D_A.

PROOF. The proof is similar to that of Kakumanu [13]. The details may be found in Doshi [4].

Acknowledgment. This paper is based on author's dissertation at Cornell University. The author is grateful to Professor N. U. Prabhu for his constant encouragement and guidance during the course of this dissertation. The author is also grateful to the referee for some very helpful suggestions.

REFERENCES

[1] BLACKWELL, DAVID (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.

[2] BLACKWELL, DAVID (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

[3] DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

[4] DOSHI, B. T. (1974). Continuous time control of Markov processes on an arbitrary state space. Ph. D. thesis, Cornell University. Also Tech. Sum. Report 1468. Math. Res. Ctr., Madison, Wisconsin.

[5] DYNKIN, E. B. (1965). *Markov Processes-I.* Springer-Verlag, Berlin.

[6] FELLER, W. (1966). *Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

[7] FLEMING, W. H. (1963). Some Markovian optimization problems. *J. Math. Mech.* **12** 131–140.

[8] FLEMING, W. H. (1969). Optimal continuous-parameter stochastic control. *SIAM Rev.* **11** 470–509.

[9] HINDERER, K. (1970). *Foundations of Nonstationary Dynamic Programming with Discrete Time Parameter.* Springer-Verlag, Berlin.

[10] HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes.* Wiley, New York.

[11] HOWARD, R. A. (1971). *Dynamic Probabilistic Systems*, **1–2**. Wiley, New York.

[12] KAKUMANU, P. K. (1969). Continuous time Markov decision models with applications to optimization problems. Tech. Report 63, Dept. of Operations Research, Cornell Univ.

[13] KAKUMANU, P. K. (1971). Continuously discounted Markov decision model with countable state and action spaces. *Ann. Math. Statist.* **42** 919–926.

[14] KUSHNER, H. J. (1966). Sufficient conditions for the optimality of a stochastic control. *SIAM J. Control* **3** 499–508.

[15] KUSHNER, H. J. (1967). Optimal discounted stochastic control for diffusion processes. *SIAM J. Control* **5** 520–531.

[16] MAITRA, A. (1968). Discounted dynamic programming on compact metric spaces. *Sankhyā Ser. A* **30** 211-216.

[17] MANDL, P. (1968). *Analytical Treatment of One-dimensional Markov Processes.* Academy of Sciences, Prague.

[18] MILLER, B. L. (1968). Finite state continuous time Markov decision processes with finite planning horizon. *SIAM J. Control* **6** 266-280.

[19] MILLER, B. L. (1968). Finite state continuous time Markov decision processes with an infinite planning horizon. *J. Math. Anal. Appl.* **22** 552-569.

[20] PLISKA, S. (1973). Single person controlled diffusions with discounted costs. *J. Optimization Theory Appl.* **12** 248-256.

[21] PLISKA, S. (1974). Controlled jump processes. Working Paper. Dept. of Ind. Engrg. and Management Sci., Northwestern Univ.

[22] PRABHU, N. U. (1965). *Stochastic Processes: Basic Theory and Applications.* Macmillan, New York.

[23] ROSS, S. M. (1968). Arbitrary state Markovian decision processes. *Ann. Math. Statist.* **39** 2118-2122.

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

[25] SCHAL, MANFRED (1972). Dynamische Optimierung unter Stetigkeits-und Kompaktheitsbedingungen. Habilitationsschrift, Univ. Hamburg.

[26] STONE, L. D. (1973). Necessary and sufficient conditions for optimal control of semi-Markov jump processes. *SIAM J. Control* **11** 187-201.

[27] STRAUCH, R. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37** 871-890.

[28] YOSHIDA, K. (1971). *Functional Analysis.* Springer-Verlag, New York.

STATISTICS CENTER
RUTGERS UNIVERSITY
NEW BRUNSWICK, NEW JERSEY 08903