COUNTABLE STATE DISCOUNTED MARKOVIAN DECISION PROCESSES WITH UNBOUNDED REWARDS

BY

JOHN M. HARRISON

TECHNICAL REPORT NO. 128
MARCH 12, 1970

SUPPORTED BY THE ARMY, NAVY AND AIR FORCE
UNDER CONTRACT Nonr-225(53)(NR-042-002)
WITH THE OFFICE OF NAVAL RESEARCH

Reproduction in Whole or in Part is Permitted for any purpose of the United States Government
Distribution of this Document is Unlimited

DEPARTMENT OF OPERATIONS RESEARCH
AND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
COUNTABLE STATE DISCOUNTED MARKOVIAN DECISION

PROCESSES WITH UNBOUNDED REWARDS

by

JOHN M. HARRISON

TECHNICAL REPORT NO. 128
March 12, 1970

Supported by the Army, Navy and Air Force under
Contract Nonr-225(53) (NR-042-002)
with the Office of Naval Research

Gerald J. Lieberman, Project Director

This research was partially supported
by the Office of Naval Research under
Contract Nonr-225(89) (NR-047-061) and
National Science Foundation Grant GK-2925

Reproduction in Whole or in Part is Permitted for
any purpose of the United States Government

Distribution of this Document is Unlimited

DEPARTMENT OF OPERATIONS RESEARCH

and

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
ABSTRACT

Countable state, finite action Markov decision processes are investigated under a criterion of maximizing expected discounted rewards over an infinite planning horizon. Well-known results of Maitra and Blackwell are generalized, their assumption of bounded rewards being replaced by the following weaker condition: the expected absolute reward to be received at time \( n+1 \) minus the actual absolute reward received at time \( n \) (as a function of the state of the system at time \( n \), the action taken at time \( n \), and the decision rule to be followed at time \( n+1 \)) can be bounded above. Under this condition it is shown that the expected discounted reward (over the infinite planning horizon) from each policy is finite and that there exists a stationary policy which is optimal. Additional results are presented concerning the policy improvement and successive approximations algorithms for computation of optimal policies. All of these results are extended to Markov renewal decision processes under one additional condition on the transition time distributions. As in Blackwell's work on discounted dynamic programming, a central role is played by Banach's fixed point theorem for contraction mappings. Examples are presented of inventory and queueing control problems which satisfy our assumptions but do not exhibit bounded rewards.
COUNTABLE STATE DISCOUNTED MARKOVIAN DECISION

PROCESSES WITH UNBOUNDED REWARDS

by

John M. Harrison

1. Introduction

Markov decision processes, first identified by Bellman [1], constitute an important special class of dynamic programs and have received considerable attention in the literature of applied mathematics in recent years. This paper deals with stationary Markov decision processes over a discrete (countable) state space $S$ where the objective is to maximize expected discounted rewards over an infinite planning horizon. Our initial concern is with discrete-time processes, and subsequently we investigate Markov renewal decision processes, the added feature of the latter being that the times between transition points (decision points) are taken to be random variables.

The discrete-time decision process with discounted objective was first studied by Howard [6], who developed an algorithm for the computation of an optimal stationary policy in the finite state and action case. Blackwell [2] then proved that an optimal stationary policy is optimal among all policies. Later papers by Maitra [9] and Blackwell [3] showed that if rewards are bounded then there exists a stationary optimal policy in the case of countable state space and finite action sets. Blackwell also proved the validity of Howard's improvement routine in this more general setting and showed that the method of successive approximations in function space converges geometrically to the optimal...
return function. This later work of course constitutes a very significant
extension of Howard's original treatment, but still the assumption of
bounded rewards is quite restrictive. Many stochastic control problems,
notably those arising in conjunction with queueing and inventory systems,
are most naturally formulated with a countable state space and unbounded
rewards. (For example, one component of the reward function may be
proportional to the state variable.) Section 2 below discusses discrete
time processes with countable state space and finite action sets when
the assumption of bounded rewards is replaced by the following weaker
condition: the expected absolute reward to be received at time \( n+1 \)
minus the actual absolute reward received at time \( n \) (as a function of
the state of the system at time \( n \), the action taken at time \( n \), and
the decision rule to be followed at time \( n+1 \)) can be bounded above.
Under this condition it is shown that the expected discounted reward
from every policy is finite and that all of Blackwell's results remain
valid.

In Blackwell's elegant treatment of the bounded rewards case, a
central role was played by the Banach fixed point theorem for contractions.
An account of the broader applicability of contraction theory in dynamic
programming was subsequently given by Denardo [4]. Contraction theory
is also central to the development of section 2 below, and the basic
structure of the arguments presented there owes much to the work of
Blackwell and Denardo. That structure is roughly as follows. First we
show that the expected discounted return from every policy is finite,
and we develop bounds that are satisfied by all such return functions.
Using these bounds, we define a set \( B \) of functions \( S \rightarrow \mathbb{R} \) (the real
line) which contains the return function for each policy. (In the bounded rewards case treated by Blackwell and Denardo, \( B \) is taken as the set of all bounded functions \( S \to R \).) We then construct a metric \( \rho \) on \( B \times B \) in such a way that \((B, \rho)\) is a complete metric space and the optimal return operator \( T \) is a contraction on \((B, \rho)\). (Blackwell and Denardo take \( \rho \) to be the metric of uniform convergence.) Application of Banach's theorem then establishes that \( T \) has a unique fixed point in \( B \), from which the desired results follow.

The Markov renewal decision process (or Markov renewal program), first formulated by Jewell [7], constitutes a very natural extension of Howard's discrete-time model. Denardo [4] showed that with bounded rewards and continuous discounting, Blackwell's results extend routinely to the Markov renewal case if a sufficiently strong regularity condition is imposed upon the transition time distributions. Our results for discrete-time decision processes with unbounded rewards also extend directly to the Markov renewal case if Denardo's regularity condition is imposed. This extension is presented in section 3 below.

In section 4 we return to discrete-time decision processes. Using slightly different assumptions (which are neither stronger nor weaker than those of section 2), it is again shown that a stationary optimal policy exists, that Howard's improvement routine is valid, and that the method of successive approximations converges to the optimal return function. Under the new assumptions, however, a somewhat stronger statement is made about the rate of convergence for successive approximations. (On the other hand, a disadvantage of the new assumptions is that none of our section 4 results extend directly to the Markov
renewal case.) The key assumption is again a first moment condition, involving both the transition probabilities and the reward structure. The arguments are slightly simpler than those of section 2, but their basic nature is very similar.

Finally, section 5 contains some remarks on the assumptions used in sections 2-4. In particular, examples are presented of some standard problems in the control of queueing and inventory systems which satisfy our assumptions but lie outside the domain of Blackwell's formulation.

2. Discrete-Time Decision Processes with Discounting

We wish to examine a discrete-time Markov decision process whose state space is some countable set S. A system is observed at each of a sequence of points in time labelled 1, 2, ... . Each time that the system is observed to be in state s ∈ S, an action "a" is chosen from a finite set A_s of possible actions and a reward r(s,a) is received. The conditional probability that the system will be observed in state t ∈ S at time n+1, given that it is observed in state s ∈ S at time n and that action a ∈ A_s is selected, and given all other previous history of the decision process, is denoted p(t|s,a). This "transition probability" function p(·|·,·) is non-negative and satisfies

\[ \sum_{t \in S} p(t|s,a) \leq 1 \quad \text{for all } s \in S \text{ and } a \in A_s. \]

For each s ∈ S and a ∈ A_s, \[ 1 - \sum_{t \in S} p(t|s,a) \] is interpreted as the probability that the process stops. Once the process has stopped, it remains stopped and earns no rewards thereafter.

We define a decision rule to be a function f which assigns to
each state \( s \in S \) a corresponding action \( f(s) \in A_s \), and we let \( F \) denote the (clearly countable) set of all possible decision rules. Thus
\[
F = \times_{s \in S} A_s.
\]
A sequence \( \pi = (f_1, f_2, \ldots) \) of decision rules (\( f_n \in F \) for all \( n = 1, 2, \ldots \)) is called a policy, and we let \( F^\infty \) denote the (clearly countable) set of all possible policies. A policy will be called stationary if all of its component decision rules are identical, and we let \( f^\infty \) denote the stationary policy all of whose components are \( f \). If \( \pi = (f_1, f_2, \ldots) \in F^\infty \) and \( g \in F \), then we will use the notation \( (g, \pi) = (g, f_1, f_2, \ldots) \in F^\infty \). The \( n \)th component of a policy is interpreted as a decision rule to be employed at time point \( n \), and hence \( (g, \pi) \) is a policy which uses \( g \) at the first time point and then uses the component rules of \( \pi \) in their given order at subsequent time points.

We will use the term "vector" throughout to mean a function \( V: S \to \mathbb{R} \), and "the \( s \)th component" of a vector \( V \) refers to \( V(s) \). Similarly a "matrix" is a function \( P: S \times S \to \mathbb{R} \) and "the \( (s, t) \)th element" of a matrix \( P \) refers to \( P(s, t) \). All vectors should be envisioned as column vectors, and we define matrix multiplication and matrix-vector multiplication in the usual way if all of the sums involved are absolutely convergent, in which case we say that the matrix product or matrix-vector product is "well defined". Convergence of vector sequences is defined to be pointwise. We will occasionally use \( 0 \) and \( 1 \) to denote a vector of zeros and a vector of ones, and the meaning of the symbols should always be clear from context. For a vector \( V \) we will use the notation \( |V| \) to denote a vector whose \( s \)th component is \( |V(s)| \), and we let \( ||V|| = \sup_{s \in S} |V(s)| \). For a matrix \( P \) we let \( ||P|| = \sup_{s \in S} \sum_{t \in S} |P(s, t)| \). We will use the obvious inequality \( ||PV|| \leq ||P|| \cdot ||V|| \).
For each $f \in F$, define $r(f)$ to be the vector which has $r(s, f(s))$ as its $s^{th}$ component, and define $P(f)$ to be the matrix which has $p(t|s, f(s))$ as its $(s, t)^{th}$ element. In order to avoid unnecessary clumsiness, we will state our two additional assumptions and prove some of their consequences before developing any further notation or even defining the objective of the "decision-maker".

Assumptions:

(i) $\sum_{t \in S} p(t|s, a)r(t, f(t))$ is absolutely convergent for each $s \in S$, $a \in A_s$, and $f \in F$.

(ii) There exists a bound $d > 0$ such that $\sum_{t \in S} p(t|s, a)|r(t, g(t))| \leq |r(s, a)| + d$ for all $s \in S$, $a \in A_s$ and $g \in F$ (or equivalently $P(f)|r(g)| \leq |r(f)| + d \cdot 1$ for all $f, g \in F$).

Assumption (i) is of course very mild, requiring only that the vector of expected rewards $P(f)r(g)$ be well defined for each pair of decision rules $f, g \in F$. Assumption (ii) is much stronger and provides the key to our analysis. Note however that (ii) is immediately implied by an assumption of bounded rewards.

Now for each policy $\pi = (f_1, f_2, \ldots) \in F^\infty$, let $P^0(\pi) = I$, the identity matrix, and let $P^n(\pi) = P(f_1)P(f_2)\cdots P(f_n)$ for $n = 1, 2, \ldots$ Thus $P^n(\pi)$ is the $n$-step transition matrix under policy $\pi$.

Lemma 1: For any policy $\pi = (f_1, f_2, \ldots) \in F^\infty$, $P^n(\pi)r(f_{n+1})$ is well defined and satisfies $|P^n(\pi)r(f_{n+1})| \leq P^n(\pi)|r(f_{n+1})| \leq |r(f_1)| + nd \cdot 1$ for $n = 0, 1, 2, \ldots$.

Proof: The proposition is trivially true for $n = 0$. Inductively,
assume it for \( n-1 \). Then using (ii) and the inductive hypothesis we have

\[
|P_n(\pi)(f_{n+1})| \leq P_n(\pi)|r(f_{n+1})| = P_{n-1}(\pi)P(\pi_n)|r(f_{n+1})| \\
\leq P_{n-1}(\pi)[|r(f_{n})| + d \cdot 1] \leq |r(f_{1})| + (n-1)d \cdot 1 + d \cdot 1 = |r(f_{1})| + nd \cdot 1.
\]

Q.E.D.

Given Lemma 1, we may meaningfully define for each \( \pi = (f_{1}, f_{2}, \ldots) \), each \( N = 1, 2, \ldots \), and each \( 0 \leq \beta < 1 \)

\[
\nu_{\beta}^{N}(\pi) = \sum_{n=0}^{N-1} \beta^{n}P_{n}(\pi)r(f_{n+1}).
\]

The \( s \)th component of \( \nu_{\beta}^{N}(\pi) \) is the expected discounted \( N \)-period reward under policy \( \pi \) when the discount factor is \( \beta \) and the system is initially in state \( s \). Our next result is that the vector of expected discounted rewards over an infinite planning horizon, \( \lim_{N \to \infty} \nu_{\beta}^{N}(\pi) \), exists and is finite for each policy \( \pi \).

**Lemma 2:** For each policy \( \pi = (f_{1}, f_{2}, \ldots) \in \Phi^{\infty} \), \( \nu_{\beta}(\pi) = \sum_{n=0}^{\infty} \beta^{n}P_{n}(\pi)r(f_{n+1}) \) is absolutely convergent and satisfies

\[
|\nu_{\beta}(\pi)| \leq (1-\beta^{-1})|r(f_{1})| + \beta d(1-\beta)^{-2}.1.
\]

**Proof:**

\[
|\nu_{\beta}(\pi)| \leq \sum_{n=0}^{\infty} \beta^{n}P_{n}(\pi)|r(f_{n+1})| \\
\leq \sum_{n=0}^{\infty} \beta^{n}[|r(f_{1})| + nd \cdot 1] \text{ by Lemma 1} \\
= (1-\beta)^{-1}|r(f_{1})| + \beta d(1-\beta)^{-2}.1.
\]

Q.E.D.

Using Lemma 2 we can now complete our formulation. We suppose that the decision maker's objective is to find a policy \( \pi^{*} \), if one exists, such that
\[ V_\beta(\pi^*) \geq V_\beta(\pi) \text{ for all } \pi \in \mathbb{F}^\infty, \]

where the discount factor \( \beta \) is fixed and satisfies \( 0 \leq \beta < 1 \). Such a policy \( \pi^* \) is called optimal. We will show that there exists a stationary optimal policy.

Let \( f_u \in \mathbb{F} \) be such that

\[ |r(f)| \leq |r(f_u)| \text{ for all } f \in \mathbb{F}. \]

The existence of such a rule is guaranteed by the fact that \( \mathbb{F} \) is defined as the Cartesian product of the finite sets \( A_s \). Further let

\[ B \ni \text{the set of all vectors } V: S \to \mathbb{R} \text{ satisfying } \]

\[ |V| \leq (1-\beta)^{-1}|r(f_u)| + \beta d(1-\beta)^{-2}\cdot 1. \]

From Lemma 2 it is clear that \( V_\beta(\pi) \in B \) for all \( \pi \in \mathbb{F}^\infty \).

**Lemma 3:** If \( \pi \in B \) and \( \pi = (f_1, f_2, \ldots) \in \mathbb{F}^\infty \) then \( \beta^n P^n(\pi)V \xrightarrow{n \to \infty} 0 \).

**Proof:** Using the definition of \( B \),

\[ |\beta^n P^n(\pi)V| \leq \beta^n P^n(\pi)|V| \]

\[ \leq \beta^n P^n(\pi)[(1-\beta)^{-1}|r(f_u)| + \beta d(1-\beta)^{-2}\cdot 1] \]

\[ \leq \beta^n (1-\beta)^{-1}[|r(f_u)| + nd \cdot 1 + \beta d(1-\beta)^{-1}\cdot 1] \xrightarrow{n \to \infty} 0 \]

by Lemma 1.

Q.E.D.

Now for each \( f \in \mathbb{F} \) define a mapping \( L_f \) on \( B \) by
\[ L_f V = r(f) + \beta P(f)V, \quad \text{VeB}. \]

That \( L_f V \) is well defined for each \( f \in F \) and \( \text{VeB} \) follows immediately from Assumption (i) and the definition of \( B \). Since \( P(f) \geq 0 \), \( L_f \) is clearly monotone.

**Lemma 4:** For each \( f \in F \), \( L_f \) maps \( B \) into itself.

**Proof:** If \( \text{VeB} \), then using Assumption (ii) we have

\[
|L_f V| = |r(f) + \beta P(f)V| \leq |r(f)| + \beta |P(f)V|
\]

\[
\leq |r(f)| + \beta (1 - \beta)^{-1} P(f) |r(f_u)| + \beta^2 d (1 - \beta)^{-2} P(f)_1
\]

\[
\leq |r(f)| + \beta (1 - \beta)^{-1} [|r(f)| + d_1] + \beta^2 d (1 - \beta)^{-2} \cdot 1
\]

\[
= (1 - \beta)^{-1} |r(f)| + \beta d (1 - \beta)^{-2} \cdot 1
\]

\[
\leq (1 - \beta)^{-1} |r(f_u)| + \beta d (1 - \beta)^{-2} \cdot 1
\]

and hence \( L_f \text{VeB} \).

Q.E.D.

The \( s^{th} \) component of \( L_f V \) is the expected discounted reward when the system begins in state \( s \), action \( f(s) \) is taken, and a terminal reward \( V(t) \) is to be received one period later if the system evolves to state \( t \). Hence Lemma 4 says that if a decision rule \( f \) is employed at the initial decision point and a terminal reward given by some vector \( \text{VeB} \) is to be received one period later, then the corresponding vector of total expected discounted rewards must also be in \( B \). It follows inductively that \( L_f^n V \) is the vector of expected discounted rewards
associated with following policy \( \tau^{\infty} \) for \( n \) periods and receiving terminal reward \( V \) in period \( n+1 \), and if \( V \in B \) then \( L^n_{\tau} V \in B \) as well.

Now to metricize the space \( B \) define \( \varphi: S \rightarrow \mathbb{R}^+ \) by

\[
\varphi(s) = \left| r(s, \tau_u(s)) \right|, \quad s \in S,
\]

and for each \( \alpha \) satisfying \( \beta < \alpha < 1 \) define \( \rho_{\alpha}: B \times B \rightarrow \mathbb{R}^+ \) by

\[
\rho_{\alpha}(U, V) = \sup_{s \in S} \frac{|U(s) - V(s)|}{\varphi(s) \cdot M_{\alpha}}, \quad U, V \in B
\]

where

\[
M_{\alpha} = \beta d(\alpha - \beta)^{-1} > 0 \quad (\beta < \alpha < 1).
\]

It is immediate from the definition of \( B \) that \( \rho_{\alpha}(U, V) < \infty \) for \( U, V \in B \). Moreover it is clear that \( \rho_{\alpha}(U, V) = 0 \) if and only if \( U = V \) and that \( \rho_{\alpha} \) is symmetric and satisfies the triangle inequality. Hence \((B, \rho_{\alpha})\) is a metric space for each \( \beta < \alpha < 1 \), and one may easily verify that it is complete. So we have

**Lemma 5:** For each \( \alpha \) satisfying \( \beta < \alpha < 1 \), \((B, \rho_{\alpha})\) is a complete metric space.

**Theorem 6:** For each \( f \in F \) and each \( \alpha \) satisfying \( \beta < \alpha < 1 \), \( L_f \) is a contraction of modulus \( \alpha \) on \((B, \rho_{\alpha})\), i.e., if \( U, V \in B \) then

\[
\rho_{\alpha}(L_f U, L_f V) \leq \alpha \rho_{\alpha}(U, V).
\]

**Proof:** Let \( L_f V(s) \) denote the \( s \)th component of \( L_f V \). We will show that for \( U, V \in B \) and \( s \in S \),

\[
[\varphi(s) + M_{\alpha}]^{-1} |L_f U(s) - L_f V(s)| \leq \alpha \rho_{\alpha}(U, V),
\]

10
which establishes the desired result.

\[
\left[\varphi(s)+M_\alpha\right]^{-1}\left|L_f U(s)-L_f V(s)\right|
\]

\[
= \left[\varphi(s)+M_\alpha\right]^{-1}\beta \sum_{t \in S} p(t|s,f(s))[U(t)-V(t)]
\]

\[
\leq \left[\varphi(s)+M_\alpha\right]^{-1}\beta \sum_{t \in S} p(t|s,f(s))[U(t)-V(t)]
\]

\[
= \left[\varphi(s)+M_\alpha\right]^{-1}\beta \sum_{t \in S} p(t|s,f(s)) \frac{|U(t)-V(t)|}{\varphi(t)+M_\alpha} [\varphi(t)+M_\alpha]
\]

\[
\leq \left[\varphi(s)+M_\alpha\right]^{-1}\beta \rho_\alpha(U,V) \sum_{t \in S} p(t|s,f(s))[M_\alpha+\sum_{t \in S} p(t|s,f(s))\varphi(t)]
\]

But

\[
\sum_{t \in S} p(t|s,f(s))\varphi(t) = \sum_{t \in S} p(t|s,f(s))|r(t,f_u(t))|
\]

\[
\leq |r(s,f(s))| + d \quad \text{by (ii)}
\]

\[
\leq |r(s,f_u(s))| + d = \varphi(s) + d
\]

and so

\[
\left[\varphi(s)+M_\alpha\right]^{-1}\left|L_f U(s)-L_f V(s)\right| \leq \left[\varphi(s)+M_\alpha\right]^{-1}\beta \rho_\alpha(U,V)[M_\alpha+\varphi(s)+d]
\]

\[
\leq \beta \rho_\alpha(U,V)[1+dM_\alpha^{-1}] = \alpha \rho_\alpha(U,V)
\]

Q.E.D.

**Corollary 7:**

a) \(L_f\) has a unique fixed point \(V_f \in B\), i.e., there exists a unique vector \(V_f \in B\) such that \(L_f V_f = V_f\).

b) \(L^n_f V_f \longrightarrow V_f\) and \(\alpha \rho_\alpha(L^n_f V_f, V_f) \leq \alpha^n (1-\alpha)^{-1} \rho_\alpha(L_f V, V)\) for all \(V \in B\) and \(\alpha\) satisfying \(\beta < \alpha < 1\).

c) \(V_f = V_\beta(r^\infty)\).
Proof: (a) and (b) follow directly from Lemma 5, Theorem 6, and the Banach fixed point theorem for contraction mappings (see [8]). Since \( L^n f \leq V^n f \) and \( \beta^n \leq \beta \), we have that \( L^n f \rightarrow V f \) as \( n \rightarrow \infty \) for all \( f \in B \), by Lemma 3. Hence \( V f = V^n f \) by (b), establishing (c).

Q.E.D.

Next, we define the familiar optimal return operator \( T \) on \( B \) by

\[
T = \max_{f \in F} L f, \quad V \in B.
\]

This vector maximum exists of course because \( F \) is defined as a Cartesian product and hence the maximization can proceed componentwise. Since \( A \) is assumed finite for all \( s \in S \), there exists for each \( V \in B \) a rule \( f \in F \) such that \( TV = L f V \). From Lemma 3 then it is clear that \( T \) maps \( B \) into itself. We now show in the standard way that \( T \) inherits the contraction property of the mappings \( L f \).

Theorem 8: For each \( \alpha \) satisfying \( \beta < \alpha < 1 \), \( T \) is a contraction of modulus \( \alpha \) on \( (B, \rho_\alpha) \).

Proof: Let \( U, V \in B \) and choose \( f, g \in F \) such that \( T f U = L f U \) and \( T g V = L g V \). Then \( TV \geq L f V \), implying \( T U - T V \leq L f U - L f V \) and hence

\[
\sup_{s \in S} \frac{T V(s) - T U(s)}{\phi(s) + M \alpha} \leq \sup_{s \in S} \frac{L f U(s) - L f V(s)}{\phi(s) + M \alpha} \leq \alpha \rho \alpha \alpha(U, V), \quad \text{using Theorem 5.}
\]

Similarly

\[
\sup_{s \in S} \frac{TV(s) - TU(s)}{\phi(s) + M \alpha} \leq \sup_{s \in S} \frac{LV(s) - LU(s)}{\phi(s) + M \alpha} \leq \alpha \rho \alpha \alpha(U, V), \quad \text{so}
\]

\[
\rho \alpha \alpha(U, V) = \sup_{s \in S} \left| \frac{T U(s) - T V(s)}{\phi(s) + M \alpha} \right| \leq \alpha \rho \alpha \alpha(U, V)
\]

Q.E.D.
From the Banach fixed point theorem then we immediately obtain the following result.

**Corollary 9:**

a) $T$ has a unique fixed point $V^* \in B$.

b) $T^n V \rightarrow V^*$ and $\rho_\alpha(T^n V, V^*) \leq \alpha^{n(1-\alpha)^{-1}} \rho_\alpha(TV, V)$ for all $V \in B$ and $\alpha$ satisfying $\beta < \alpha < 1$.

Our final result (a) proves the validity of Howard's improvement routine in this more general setting, (b) provides a necessary and sufficient condition for optimality of a policy, and (c) proves the existence of a stationary optimal policy.

**Theorem 10:**

a) If $\nu_\beta(g, \pi) > \nu_\beta(\pi)$, then $\nu_\beta(g^\infty) > \nu_\beta(\pi)$.

b) A policy $\pi$ is optimal if and only if $\nu_\beta(\pi) \geq \nu_\beta(g, \pi)$ for all $g \in F$.

c) There exists a stationary optimal policy $f^\infty$ and $\nu_\beta(f^\infty) = V^*$.

**Proof:**

a) If $L_g \nu_\beta(\pi) = \nu_\beta(g, \pi) > \nu_\beta(\pi)$ then (since $L_g$ is monotone)

$L^n_g \nu_\beta(\pi) = L^n_g \nu_\beta(\pi) > \nu_\beta(\pi)$ for all $n \geq 1$. Hence $\nu_\beta(g^\infty) > L_g \nu_\beta(\pi) > \nu_\beta(\pi)$ by Corollary 7.

b) The "only if" part follows from (a) and the fact that either $\nu_\beta(\pi) \geq \nu_\beta(g, \pi)$ for all $g \in F$ or else $\nu_\beta(\pi) < \nu_\beta(g, \pi)$ for some $g \in F$. Conversely, if $\nu_\beta(\pi) \geq L_g \nu_\beta(\pi)$ for all $g \in F$, let $\pi' = (f'_1, f'_2, \ldots)$ be any other policy. Then (by the monotonicity of the operators $L_f$)

$L_{f_1} L_{f_2} \cdots L_{f_n} \nu_\beta(\pi) \leq \nu_\beta(\pi)$ for all $n$, or $\nu_{\beta(\pi')} + \beta^n P^n(\pi') \nu_\beta(\pi) \leq \nu_\beta(\pi)$.
for all \( n \). Letting \( n \to \infty \) and invoking Lemma 3, we have \( V_\beta(\pi') \leq V_\beta(\pi) \). Hence \( \pi \) is optimal.

c) There exists \( f \in F \) such that \( TV^* = L_\pi V^* = V^* \). From Corollary 7 then \( V^* = V_\beta(f^\infty) \). Moreover \( V_\beta(f^\infty) = V^* = TV^* \geq L_\pi V^* \) for all \( g \in F \), so \( f^\infty \) is optimal by (b).

Q.E.D.

Corollary 9(a) establishes that the extremal equation

\[
V = \max_{f \in F} \left[ r(f) + \beta P(f)V \right]
\]

has a unique solution \( V^* \in B \), and Theorem 10(c) shows that \( V^* \) is the vector of expected discounted returns from an optimal policy (often called the "optimal return function"). Corollary 9(b) shows that \( V^* \) can be computationally approximated by repeated application of the optimal return operator \( T \) to any arbitrary initial vector \( V \in B \). This computational procedure is known variously as "value iteration" (see Howard [6]) and "the method of successive approximations" (see Bellman [1]).

3. The Extension to Markov Renewal Decision Processes

We now wish to expand our formulation to allow for the possibility of transition times which are random variables. Specifically, the changes required in the previous formulation are as follows. Suppose that a transition occurs at time \( t \geq 0 \), that the system is then observed to be in state \( s \in S \), and that an action \( a \in A_s \) is selected. Then a reward \( r(s,a) \) is received (at time \( t \)), and the joint (conditional) probability that the next transition is to state \( s' \in S \) and
occurs no later than at time $t+\tau$ is given by $Q(s', \tau | s, a)$, where

a) for each $s', s \in S$ and $a \in A$, $Q(s', \cdot | s, a)$ is real-valued, non-decreasing, right-continuous and satisfies $Q(s', \tau | s, a) = 0$ for all $\tau < 0$;

b) for each $s \in S$ and $a \in A$, $\sum_{s' \in S} Q(s', \cdot | s, a) \leq 1$.

As before, for each $s \in S$ and $a \in A$, we interpret $1 - \sum_{s' \in S} Q(s', \cdot | s, a)$ as the probability that the next transition is to some external stopping state.

We assume that decisions can be made (actions taken) only at the times of transition. Decision rules and policies are defined as before, where now the $n^{th}$ component of a policy is interpreted as a decision rule to be employed at the $n^{th}$ transition point, regardless of when it may occur in time. By convention, we will assume that the first "transition" occurs at time zero, the initial state of the system being determined in this way. The final change in our formulation is that now we assume rewards to be discounted continuously in time using some positive interest rate $\gamma > 0$, so that a reward of $x$ received at time $t$ is equivalent in value to a reward of $xe^{-\gamma t}$ received at time zero.

For each $s, s' \in S$ and each $a \in A$, define $p(s' | s, a) = Q(s', \cdot | s, a)$.

For each $f \in F$ let $r(f)$ be as before and let $P(f)$ be defined from $p(\cdot | \cdot, \cdot)$ as before. Similarly, for each $\pi = (f_1, f_2, \ldots) \in F^\infty$ define $P^n(\pi), n = 0, 1, 2, \ldots$ in terms of the matrices $P(f_n)$ as before. Clearly $r(f)$, $P(f)$ and $P^n(f)$ play the same role in Markov renewal decision processes as did the corresponding quantities in the discrete-time version.
Now for each $s, s' \in S$ and $a \in A_s$ we define
\[
\beta(s, s', a) = \begin{cases} 
0 & \text{if } p(s' | s, a) = 0, \\
p(s' | s, a)^{-1} \int_0^\infty e^{-\gamma t} dQ(s', \tau | s, a) & \text{if } p(s' | s, a) > 0.
\end{cases}
\]

The quantity $\beta(s, s', a)$ is the expected discount factor to be applied to rewards earned at the next transition point, given that the system is currently in state $s$, action $a$ is taken, and the next transition is to state $s'$. Clearly, $0 \leq \beta(s, s', a) \leq 1$ and $\beta(s, s', a) = 1$ if and only if $Q(s', \cdot | s, a)$ places all of its mass at the origin. The following assumption, originally used by Denardo [4], rules out the latter possibility. In fact it further guarantees that only finitely many transitions can occur in a finite interval of time (as will be shown in section 5 below), and hence, using the terminology of Pyke [9] somewhat loosely, we refer to it as a "regularity" condition.

**Assumption (iii):** There exists $\beta < 1$ such that $\beta(s', s, a) \leq \beta$ for all $s, s' \in S$ and $a \in A_s$.

Now for each $f \in F$ define $\hat{P}(f)$ to be a matrix which has $p(s' | s, a)\beta(s, s', a)$ as its $(s,s')$th element. For each policy $\pi = (f_1, f_2, \ldots) \in F^\omega$ define $\hat{P}^0(\pi) = I$ and $\hat{P}^n(\pi) = \hat{P}(f_1)\hat{P}(f_2)\ldots\hat{P}(f_n)$ for $n = 1, 2, \ldots$. Clearly by (iii)
\[
\hat{P}(f) \leq \beta \hat{P}(f) \quad \text{for all } f \in F,
\]
\[
\hat{P}^n(\pi) \leq \beta^n \hat{P}^n(\pi) \quad \text{for all } f \in F^\omega \text{ and } n = 0, 1, 2, \ldots.
\]

The substochastic matrix $\hat{P}(f)$ plays the same role in Markov renewal
decision processes as does the matrix $\beta P(f)$ in discrete-time processes.

As an analog to the vector $v_{\beta}^{N}(\pi)$ defined in section 2 for discrete-time processes we next define

$$v_{\gamma}^{N}(\pi) = \sum_{n=0}^{N-1} P^{n}(\pi)r(f_{n+1}) \quad (\pi = (f_1, f_2, \ldots) \in F^{\infty}, \quad N = 1, 2, \ldots).$$

As was shown by Jewell [7], the $s^{th}$ component of $v_{\gamma}^{N}(\pi)$ is the expected discounted reward earned by policy $\pi$ through $N$ transitions (including rewards earned as a consequence of the action selected at the $n^{th}$ transition point) when the initial state of the system is $s$. Similarly

$$v_{\gamma}(\pi) = \lim_{N \to \infty} v_{\gamma}^{N}(\pi) = \sum_{n=0}^{\infty} P^{n}(\pi)r(f_{n+1}) \quad (\pi = (f_1, f_2, \ldots) \in F^{\infty})$$

is the vector of expected discounted rewards under policy $\pi$ over an infinite planning horizon. (Given the "regularity" implications of assumption (iii), to be proved in section 5, we need not differentiate between an infinite time horizon and infinitely many transitions.) We seek a policy $\pi^{*}$, called optimal, such that

$$v_{\gamma}(\pi^{*}) \geq v_{\gamma}(\pi) \text{ for all } \pi \in F^{\infty}.$$  

Making assumptions (i)-(iii) and using these new definitions, we find that a precise analog is obtainable for each of the results developed in section 2. Those analogs will now be stated but not proved, as their proofs differ only trivially from the corresponding ones in section 2.

**Lemma 1**: For any policy $\pi = (f_1, f_2, \ldots) \in F^{\infty}$, $P^{n}(\pi)r(f_{n+1})$ is well
defined and satisfies \[ |P^n(\pi)r(f_{n+1})| \leq P^n(\pi)|r(f_n)| \leq |r(f_1)| + nd \cdot l \]
for \( n = 0, 1, 2, \ldots \).

**Lemma 2:** For each policy \( \pi = (f_1, f_2, \ldots) \in \mathcal{F}^\infty \), \( V_\gamma(\pi) = \sum_{n=0}^{\infty} P^n(\pi)r(f_{n+1}) \) is absolutely convergent and satisfies \( |V_\gamma(\pi)| \leq (1-\beta)^{-1}|r(f_1)| + \beta d(1-\beta)^{-2} \cdot l \).

Then with \( B \) defined exactly as before, \( \rho_\alpha \) defined exactly as before for each \( \alpha \) satisfying \( \beta < \alpha < 1 \), and \( L_f \) and \( T \) defined by

\[
L_f V = r(f) + \mathcal{P}(f)V \quad (f \in \mathcal{F}, V \in B),
\]

\[
T V = \max_{f \in \mathcal{F}} L_f V \quad (V \in B),
\]

we have the following results.

**Lemma 3:** If \( V \in B \) and \( \pi = (f_1, f_2, \ldots) \in \mathcal{F}^\infty \) then \( P^n(\pi)V \xrightarrow{n \to \infty} 0 \).

**Lemma 4:** For each \( f \in \mathcal{F} \), \( L_f \) maps \( B \) into itself.

**Lemma 5:** For each \( \alpha \) satisfying \( \beta < \alpha < 1 \), \( (B, \rho_\alpha) \) is a complete metric space.

**Theorem 6:** For each \( f \in \mathcal{F} \) and each \( \alpha \) satisfying \( \beta < \alpha < 1 \), \( L_f \) is a contraction of modulus \( \alpha \) on \( (B, \rho_\alpha) \).

**Corollary 7:**

a) \( L_f \) has a unique fixed point \( V_f \in B \).
b) $L_f^n V \xrightarrow[n \to \infty]{} V_f$, and for all $V \in B$ and $\alpha$ satisfying $\beta < \alpha < 1$ we have $\rho_\alpha(L_f^n V, V_f) \leq \alpha^n (1-\alpha)^{-1} \rho_\alpha(L_f V, V)$.

c) $V_f = V_\gamma(f^\infty)$.

**Theorem 8:** For each $\alpha$ satisfying $\beta < \alpha < 1$, $T$ is a contraction of modulus $\alpha$ on $(B, \rho_\alpha)$.

**Corollary 9:**

a) $T$ has a unique fixed point $V^\ast \in B$.

b) $T^n V \xrightarrow[n \to \infty]{} V^\ast$ and $\rho_\alpha(T^n V, V^\ast) \leq \alpha^n (1-\alpha)^{-1} \rho_\alpha(T V, V)$ for all $V \in B$ and $\alpha$ satisfying $\beta < \alpha < 1$.

**Theorem 10:**

a) If $V_\gamma(g, \pi) > V_\gamma(\pi)$, then $V_\gamma(g^\infty) > V_\gamma(\pi)$.

b) A policy $\pi$ is optimal if and only if $V_\gamma(\pi) \geq V_\gamma(g, \pi)$ for all $g \in F$.

c) There exists a stationary optimal policy $f^\infty$ and $V_\gamma(f^\infty) = V^\ast$.

**h. Alternate Assumptions for Discounted-Time Processes**

Although the results obtained in section 2 and 3 are quite strong, the computational implications of Corollary 9(b) are somewhat weaker than one would like. It shows that the method of successive approximations (i.e., repeated application of the optimal return operator $T$) converges to the optimal return function $V^\ast$, but because of the way in
which we have metricized the space $B$, it does not guarantee uniform convergence. Hence one is led naturally to seek alternate and perhaps stronger assumptions which will lead to a function space on which the metric of uniform convergence can be used. Such a set of assumptions is presented in this section, and under these assumptions we are able to obtain precise analogs for each of the results in section 2. The analog to Corollary 9(b) then gives the desired uniform convergence. We emphasize that the discussion of this section is restricted to discrete-time Markov decision processes, and the basic notation and terminology of section 2 will be maintained. It is interesting that the results obtained here do not extend in any direct way to the Markov renewal case. In fact, assumptions (i)'-(iii)' below are not sufficient to guarantee that the expected discounted reward from every policy is finite in a Markov renewal setting, even if they are combined with very strong conditions on the transition time distributions. (This fact will be demonstrated with an example in section 5 below.)

Assumptions:

(i)' $\sum_{t \in S} p(t|s,a) r(t,f(t))$ is absolutely convergent for each $s \in S$, $a \in A_s$ and $f \in F$.

(ii)' There exists a bound $d > 0$ such that $|\sum_{t \in S} p(t|s,a) r(t,g(t)) - r(s,a)| \leq d$ for all $s \in S$, $a \in A_s$ and $g \in F$ (or equivalently $\|P(f)r(g) - r(f)\| \leq d$ for all $f, g \in F$).

(iii)' $\max_{a \in A_s} r(s,a) - \min_{a \in A_s} r(s,a)$ is a bounded function of $s$ over $S$.

Strictly speaking, assumptions (i)'-(iii)' are neither stronger nor
weaker than assumptions (i) and (ii) of section 2. Again it is apparent that (i)'-(iii)' are implied by an assumption of bounded rewards.

**Lemma 1:** For any policy \( \pi = (f_1, f_2, \ldots) \in \mathcal{F}^\infty \), \( P^\pi f_{n+1} \) is well defined and satisfies \( \|P^\pi f_{n+1} - f_1\| \leq nd \), \( n = 0, 1, 2, \ldots \).

**Proof:** The proposition is trivially true for \( n = 0 \). Inductively, assume it for \( n-1 \). Then

\[
\|P^\pi f_{n+1} - f_1\| = \|P^{n-1}\pi f_n f_{n+1} - f_1\| + \|P^{n-1}\pi f_n - f_1\| \leq \|P^{n-1}\pi f_n f_{n+1} - f_1\| + \|P^{n-1}\pi f_n - f_1\| \\
\leq 1 \cdot d + (n-1)d = nd ,
\]

using the inductive hypothesis and (ii)'.

Q.E.D.

**Lemma 2:** If the discount factor is \( 0 \leq \beta < 1 \), then for each policy \( \pi = (f_1, f_2, \ldots) \), \( V_\beta(\pi) = \sum_{n=0}^\infty \beta^n P^\pi f_{n+1} \) is absolutely convergent and satisfies \( \|V_\beta(\pi) - (1-\beta)^{-1} f_1\| \leq \beta d(1-\beta)^{-2} \).

**Proof:**

\[
\|V_\beta(\pi) - (1-\beta)^{-1} f_1\| = \|\sum_{n=0}^\infty \beta^n P^\pi f_{n+1} - f_1\| \leq \sum_{n=0}^\infty \beta^n \|P^\pi f_{n+1} - f_1\| \leq \sum_{n=0}^\infty \beta^n (nd) = \beta d(1-\beta)^{-2}, \text{ using Lemma 1.}
\]

Q.E.D.

Now let the decision rules \( f_L, f_U \in \mathcal{F} \) be such that

\[ r(f_L) \leq r(f) \leq r(f_U) \text{ for all } f \in \mathcal{F} , \]
and then define the space $B$ by

$$
B = \text{the set of all vectors } V: S \rightarrow R \text{ satisfying }
(l-\beta)^{-1}r(f_L)-\beta d(l-\beta)^{-2}\cdot 1 \leq V \leq (l-\beta)^{-1}r(f_U)+\beta d(l-\beta)^{-2}\cdot 1,
$$

where $1$ denotes the vector of ones.

\textbf{Lemma 3:} If $V \epsilon B$ and $\pi = (f_1, f_2, \ldots) \epsilon F^\infty$, then $\beta^n P^n(\pi) V \xrightarrow{n \to \infty} 0$.

\textbf{Proof:} By definition of $B$, it suffices to show that $\beta^n P^n(\pi) r(f_L) \xrightarrow{n \to \infty} 0$ and $\beta^n P^n(\pi) r(f_U) \xrightarrow{n \to \infty} 0$. By Lemma 1, $\|\beta^n P^n(\pi) r(f_L) - \beta^n r(f_L)\| \xrightarrow{n \to \infty} 0$, and since clearly $\beta^n r(f_L) \xrightarrow{n \to \infty} 0$, we conclude that $\beta^n P^n(\pi) r(f_L) \xrightarrow{n \to \infty} 0$. The corresponding statement for $\beta^n P^n(\pi) r(f_U)$ follows in identical fashion.

Q.E.D.

From Lemma 2 it is apparent that for each $\pi \epsilon F^\infty$ the return function $V_\beta(\pi) \epsilon B$.

\textbf{Lemma 4:} For each $f \epsilon F$, $L_f$ maps $B$ into itself.

\textbf{Proof:} If $V \epsilon B$, then $V \geq (l-\beta)^{-1}r(f_L)-\beta d(l-\beta)^{-2}\cdot 1$, and hence

$$
L_f V = r(f) + \beta P(f) V
\geq r(f) + \beta(l-\beta)^{-1}r(f)r(f_L) - \beta^2 d(l-\beta)^{-2}P(f) \cdot 1
\geq r(f) + \beta(l-\beta)^{-1}P(f) r(f_L) - \beta^2 d(l-\beta)^{-2} \cdot 1.
$$

But $P(f)r(f_L) \geq r(f) - d \cdot 1$ by (ii)', so

$$
L_f V \geq [1 + \beta(l-\beta)^{-1}]r(f) - [\beta(l-\beta)^{-1} + \beta^2 (l-\beta)^2 d] \cdot 1
$$

22
\[(1-\beta)^{-1}r(f) - \beta d(1-\beta)^{-1} \leq (1-\beta)^{-1}r(f_U) - \beta d(1-\beta)^{-1} \cdot 1.\]

That \(L_f v \leq (1-\beta)^{-1}r(f_U) + \beta d(1-\beta)^{-1} \cdot 1\) follows similarly.

Q.E.D.

Now to metricize \(B\) define \(\rho: B \times B \rightarrow \mathbb{R}^+\) by

\[
\rho(U, V) = ||U-V|| = \sup_{s \in S} |U(s) - V(s)|, \quad U, V \in B.
\]

It follows from assumption (iii)' that \(\rho(U, V) < \infty\) for all \(U, V \in B\) (since (iii)' implies \(||r(f_U) - r(f_L)|| < \infty\)), and hence the following lemma is elementary.

**Lemma 5:** \((B, \rho)\) is a complete metric space.

**Theorem 6:** For each \(f \in F\), \(L_f\) is a contraction of modulus \(\beta\) on \((B, \rho)\).

**Proof:** For any \(U, V \in B\)

\[
\rho(L_f U, L_f V) = \sup_{s \in S} |\beta \sum_{t \in S} p(t | s, f(s)) U(t) - V(t)|
\]

\[
\leq \beta \sup_{s \in S} \sum_{t \in S} p(t | s, f(s)) |U(t) - V(t)|
\]

\[
\leq \beta \sup_{s \in S} \sum_{t \in S} p(t | s, f(s)) \rho(U, V)
\]

\[
\leq \beta \rho(U, V)
\]

Q.E.D.

**Corollary 7:**

a) \(L_f\) has a unique fixed point \(V_f \in B\).

b) \(L^n_f V \rightarrow V_f\) and \(\rho(L^n_f V, V_f) \leq \alpha^n(1-\alpha)^{-1} \rho(L_f V, V)\) for all \(V \in B\).
c) \( V_f = V_\beta(f^\infty) \).

Proof: (a) and (b) follow directly from Lemma 5, Theorem 6 and the Banach fixed point theorem. Since \( L_f^n V = V_\beta^n(f^\infty) + \beta^n P^n(f^\infty)V \), we have
\[
L_f^n V \xrightarrow{n \to \infty} V_\beta(f^\infty)
\]
for all \( V \in B \) by Lemma 3. Hence \( V_f = V_\beta(f^\infty) \) by (b).

Theorem 8: The optimal return operator \( T \) is a contraction of modulus \( \beta \) on \( (B, \rho) \).

Proof: Let \( U,V \in B \) and choose \( f,g \in F \) such that \( TU = L_f U \) and \( TV = L_g V \). Then \( TV \geq L_f V \), implying \( TU - TV \leq L_f U - L_f V \) and hence
\[
\sup_{s \in S} [TU(s) - TV(s)] \leq \sup_{s \in S} [L_f U(s) - L_f V(s)] \leq \rho(U,V) \text{ by Theorem 6. Similarly, } \sup_{s \in S} [TV(s) - TU(s)] \leq \rho(V,U) = \rho(U,V). \text{ So } \rho(TU,TV) \leq \rho(U,V).
\]

Q.E.D.

Corollary 9:

a) \( T \) has a unique fixed point \( V^* \in B \).

b) \( T^n V \xrightarrow{n \to \infty} V^* \) and \( ||T^n V - V^*|| \leq \beta(1-\beta)^{-1}||TV - V|| \) for all \( V \in B \).

Theorem 10:

a) If \( V_\beta(g,\pi) > V_\beta(\pi) \), then \( V_\beta(g^\infty) > V_\beta(\pi) \).

b) A policy \( \pi \) is optimal if and only if \( V_\beta(\pi) \geq V_\beta(g,\pi) \) for all \( g \in F \).

c) There exists a stationary optimal policy \( f^\infty \) and \( V_\beta(f^\infty) = V^* \).

Proof: Identical to proof of Theorem 10, section 2.
5. **Additional Remarks**

Assumptions (i)' and (iii)' of section 4 are actually sufficient in themselves to guarantee the existence of a stationary optimal policy. It is the addition of assumption (iii)', however, which allows us to use the metric of uniform convergence over the space $B$, thus ultimately guaranteeing the uniform convergence of successive approximations.

(Incidentally, if assumption (iii)' is retained, then assumption (ii)' is clearly equivalent to the simpler requirement that $\|P(f)r(z)-r(f)\|$ be bounded over $f \in F$. ) This writer has been unable to find any condition analogous to (iii)' which will guarantee uniform convergence of successive approximations when added to assumptions (i) and (ii) of section 2.

In section 4 it was stated that assumptions (i)'-(iii)' are not sufficient to guarantee that the expected discounted reward from every policy is finite in a Markov renewal setting, even if they are combined with very strong conditions on the transition time distributions. To demonstrate that fact, consider the following Markov renewal decision process in which only one action is available in each state (so that what we actually have is just a stochastic process with rewards). The state space is

$$S = \{(I_1, \ldots, I_n) : n = 1, 2, \ldots; \ I_j = 1 \text{ or } -1 \text{ for each } j = 1, \ldots, n\},$$

and the reward associated with state $(I_1, \ldots, I_n)$ is

$$r(I_1, \ldots, I_n) = \sum_{j=1}^{n} I_j S^j.$$

The transition time distributions are such that starting from state
the probability is one-half that the system moves next to state \((I_1, \ldots, I_n, i)\) after one unit of time, and the probability is one-half that the system moves next to state \((I_1, \ldots, I_n, 1)\) after two units of time. Finally let the interest rate by \(\gamma = -\log(\frac{1}{2})\), so that \(e^{-\gamma_1} = \frac{1}{2}\) and \(e^{-\gamma_2} = \frac{1}{4}\). Assumption (iii) of section 4 is trivially satisfied by this example since only one action is available in each state, and it is clear that assumption (i)' of section 4 is satisfied as well. Assumption (ii)' is satisfied because the expected (undiscounted) reward to be received at transition point \(n+1\) is exactly equal to the actual reward received at transition point \(n\), regardless of the state at transition point \(n\). The interested reader can verify, however, that the expected discounted reward over infinitely many transitions diverges to \(+\infty\) in this example, regardless of the initial state of the system.

Pyke [10] uses the term "regular" to describe a Markov renewal process (without any decision-making aspects) in which the number of transitions in any finite length of time is finite almost everywhere. Extending this terminology to Markov renewal decision processes, we establish the following definition.

**Definition:** A Markov renewal decision process (as formulated in section 3) is called "regular" if for each \(t > 0\) the probability of infinitely many transition in \([0,t]\) is zero under each policy, \(\pi \in F^\infty\) and for each initial state of the system \(s_0 \in S\).

**Theorem 1:** If the transition time distributions \(Q(\cdot, \cdot | \cdot, \cdot)\) of a Markov renewal decision process satisfy assumption (iii) of section 3 for some interest rate \(\gamma > 0\), then it is regular.
Proof: Let \( \pi = (f_1, f_2, \ldots) \) denote the (arbitrary) policy to be followed, \( s_n \) the state of the system after the \( n \)th transition \((n = 0, 1, 2, \ldots)\), and \( \tau_n \) the time required for the \( n \)th transition \((n = 1, 2, \ldots)\). Let \( T_n = \tau_1 + \tau_2 + \cdots + \tau_n \), the total time required for the first \( n \) transitions \((n = 1, 2, \ldots)\). Let \( G_\alpha(\tau_j | s_0, \ldots, s_n) \) denote the Laplace-Stieltjes transform (L.S.T.) of the conditional distribution of \( \tau_j \) given \( s_0, s_1, \ldots, s_n \) \((n = 1, 2, \ldots; j = 1, 2, \ldots, n)\), \( \alpha \) being the dummy variable. Let \( G_\alpha(T_n | s_0, \ldots, s_n) \) denote the corresponding L.S.T. of the conditional distribution of \( T_n \) \((n = 1, 2, \ldots)\).

(In this notation and in that to follow we suppress dependence on the policy \( \pi \) for simplicity.) Defining \( \beta(\cdot, \cdot, \cdot) \) as in section 3 and letting \( \gamma > 0 \) denote the interest rate, it is clear that

\[
G_\gamma(T_n | s_0, \ldots, s_n) = \beta(s_{j-1}, s_j, f_{j-1}(s_{j-1})).
\]

Then by the independence of successive transition times,

\[
G_\gamma(T_n | s_0, \ldots, s_n) = \prod_{j=1}^{n} G_\gamma(\tau_j | s_0, \ldots, s_n) = \prod_{j=1}^{n} \beta(s_{j-1}, s_j, f_{j-1}(s_{j-1})).
\]

Thus if assumption (iii), section 3, is satisfied there exists \( 0 \leq \beta < 1 \) such that

\[
G_\gamma(T_n | s_0, \ldots, s_n) \leq \beta^n.
\]

Now letting \( G_\alpha(T_n | s_0) \) denote the L.S.T. of the conditional distribution of \( T_n \) given only the initial state of the system (with dummy variable \( \alpha \)), we have
\[ G_\gamma(T_n|s_0) = \sum_{s_1, \ldots, s_n \in S} p(s_j|s_{j-1}, f_{j-1}(s_{j-1})) G_\gamma(T_n|s_0, \ldots, s_n) \leq \beta^n. \]

So for any \( t > 0 \) and integer \( n \geq 1 \)

\[ \beta^n \geq G_\gamma(T_n|s_0) \geq e^{-\gamma t} \text{Prob}[T_n \leq t|s_0]. \]

Leaving \( t \) fixed and letting \( n \to \infty \) we obtain

\[ \text{Prob}[T_n \leq t \text{ for all } n|s_0] = \lim_{n \to \infty} \text{Prob}[T_n \leq t|s_0] \leq \lim_{n \to \infty} e^{-\gamma t} \beta^n = 0. \]

Q.E.D.

The following is a simple example of a discrete-time Markov decision process arising in inventory theory which satisfies assumptions (i) and (ii) of section 2 and assumptions (i)'-(iii)' of section 4 but does not meet Blackwell's requirement of bounded rewards. A single-product inventory system is examined (reviewed) at regular intervals. During each review period demands for the product are received from outside the system. Demand is stationary, the probability being \( p_n \) that demands for a total of \( n \) items arrive during any period. We assume that the mean demand per period, \( \mu = \sum_{n=1}^{\infty} np_n \), is finite. At each review point the decision-maker observes the number of unfilled demands for stock currently on hand, and he places an order for \( j \leq \min(s, K) \) items of stock, where \( s \) is the current unfilled demand and \( K \) is a maximum order quantity. This order is delivered at the
end of the review period, at which point \( j \) demands depart the system. The cost of ordering \( j \) items in any period is \( c_j \) \( (j = 0, 1, \ldots, K) \), and a cost of \( h \) is incurred for each item of unfilled demand held through an entire review period. A reward of \( d \) is received for each item of demand satisfied. All demands are retained in the system until they are satisfied. The problem of determining an optimal ordering policy can be formulated as a discrete-time Markov decision process where the "state of the system" is the current number of items of unfilled demand (so \( S = \{0, 1, 2, \ldots\} \)), the decision-maker's "action" is his order quantity (so \( A_s = \{0, 1, \ldots, \min(s, K)\} \) for all \( s \in S \)), the transition probabilities are given by

\[
    p(t | s, j) = \begin{cases} 
    p_{t-s+j} & \text{for } t = s-j, s-j+1, \ldots \\
    0 & \text{otherwise}
    \end{cases}
\]

and the rewards are given by

\[
    r(s, j) = -c_j + jd - (s-j)h
\]

(which is clearly unbounded below). It is easily verified that this model satisfies the assumptions of sections 2 and 4 (even if \( \mu > K \)), not only in the simple form presented here but also when additional complicating features are added (e.g., assuming that many products are stocked simultaneously, allowing random time lags in order delivery, admitting certain types of non-linear holding costs, etc.).

Finally we remark that many problems arising in the control of queueing systems can be formulated as Markov renewal decision processes.
which do not exhibit bounded rewards but satisfy assumptions (i)-(iii) of sections 2 and 3. In such problems the "state of the system" at time \( t \) is normally defined to be \( Q(t) \), the number of customers currently in the system (plus possibly some indicator of the current mode of control). Typically there exists a sequence \( \{t_n\} \) of "regeneration points" such that \( \{Q(t_n)\} \) is an "imbedded" Markov chain, and decisions are allowed only at regeneration points. One component of the cost structure is usually a customer holding cost which grows without bound as the number of customers increases (assuming that the queue length is unrestricted). Specific examples of such problems are the models studied by Heyman [5] and Yechiali [11]. The former author investigates optimal on-off control of the server in an \( \text{M|G|1} \) queue, while the latter is concerned with optimal input control in a \( \text{GI|M|1} \) queue. The interested reader may verify that these models do in fact satisfy our assumptions, even in some versions much more complicated than those considered by their respective authors.
REFERENCES

Countable State Discounted Markov Decision Processes with Unbounded Rewards

Technical Report

HARRISON, John M.

March 12, 1970

Technical Report No. 128

Distribution of this document is unlimited

Logistics and Mathematical Statistics Branch
Office of Naval Research
Washington, D. C. 20360

Countable state, finite action Markov decision processes are investigated under a criterion of maximizing expected discounted rewards over an infinite planning horizon. Well-known results of Maitra and Blackwell are generalized, their assumption of bounded rewards being replaced by the following weaker condition: the expected absolute reward to be received at time \( n+1 \) minus the actual absolute reward received at time \( n \) (as a function of the state of the system at time \( n \), the action taken at time \( n \), and the decision rule to be followed at time \( n+1 \)) can be bounded above. Under this condition it is shown that the expected discounted reward (over the infinite planning horizon) from each policy is finite and that there exists a stationary policy which is optimal. Additional results are presented concerning the policy improvement and successive approximations algorithms for computation of optimal policies. All of these results are extended to Markov renewal decision processes under one additional condition on the transition time distributions. As in Blackwell's work on discounted dynamic programming a central role is played by Banach's fixed point theorem for contraction mappings. Examples are presented of inventory and queuing control problems which satisfy our assumptions but do not exhibit bounded rewards.
### Key Words

- Contraction mapping
- Discounting
- Dynamic programming
- Fixed point
- Markov decision process
- Markov renewal decision process
- Policy improvement
- Regularity condition
- Stationary optimal policy
- Successive renewal optimal decision process

### Instructions

1. **Originating Activity:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (corporate author) issuing the report.

2. **Report Security Classification:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

3. **Group:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

4. **Report Title:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

5. **Descriptive Notes:** If appropriate, enter the type of report, e.g., interim progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

6. **Author(s):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

7. **Total Number of Pages:** The total page count should follow manual pagination procedure, i.e., enter the number of pages containing information.

8. **Number of References:** Enter the total number of references cited in the report.

9. **Contract or Grant Number:** If appropriate, enter the applicable number of the contract or grant on which the report was written.

10. **Project Number:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

11. **Originator's Report Number(s):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

12. **Availability/Limitation Notices:** Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

   1. "Qualified requesters may obtain copies of this report from DDC."
   2. "Foreign announcement and dissemination of this report by DDC is not authorized."
   3. "U.S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through ___.
   4. "U.S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through ___.
   5. "All distribution of this report is controlled. Qualified DDC users shall request through ___.

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

13. **Supplementary Notes:** Use for additional explanatory notes.

14. **Sponsoring Military Activity:** Enter the name of the sponsoring organization or laboratory sponsoring (paying for) the research and development. Include address.

15. **Abstract:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

   It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (T), (S), (C), or (U). There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

16. **Key Words:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.