MARKOVIAN DECISION PROCESSES WITH RECURSIVE REWARD FUNCTIONS

Furukawa, Nagata
Department of Statistics, University of California | Kyushu University

Iwamoto, Seiichi
Department of Mathematics, Kyushu University

https://doi.org/10.5109/13074

出版情報：統計数理研究. 15 (3/4), pp.79-91, 1973-03. Research Association of Statistical Sciences
バージョン：
権利関係：

MARKOVIAN DECISION PROCESSES WITH RECURSIVE REWARD FUNCTIONS

By

Nagata FURUKAWA* and Seiichi IWAMOTO**

(Received September 28, 1972; revised February 25, 1973)

1. Introduction.

In this paper we shall treat Markovian decision processes associated with recursive reward functions, and we shall study the existence of optimal policies in several senses and the properties of them.

Our Markovian decision processes are specified by four objects S, A, q and g. S and A are non-empty Borel sets, q is a regular conditional probability on S given $S\times A$, and g is a Baire function defined on $\bigcup_{n=2}^{\infty}\bigcup_{1\leqq m<n}[S_m\times A_m\times S_{m+1}\times A_{m+1}\times\cdots\times S_n\times R]$ $\cup\bigcup_{m=1}^{\infty}[S_m\times A_m\times S_{m+1}\times A_{m+1}\times\cdots]$, where $S_k=S$, $A_k=A$ for $k=1,2,\cdots$ and $R=(-\infty,\infty)$. We interpret S as the set of states of our system, and A as the set of actions available to us at each stage. The set of actions available is assumed to be independent of the state. When the system is at a state s and we taken an action a, the system moves to a new state s' according to the probability distribution $q(\cdot|s, a)$. The process is again repeated from the new state s', and so on. If we terminate immediately after observing the past history $h_n=(s_1, a_1, s_2, a_2, \cdots, s_n)$ of the system up to the n-th stage, we are assumed to receive a reward $g(s_1, a_1, \cdots, a_{n-1}, s_n, v(s_1, a_1, \cdots, a_{n-1}, s_n))$ (abbreviated as $g(h_n, v(h_n))$, where $v(h_n)$ is a generalized terminal reward at the past history $h_n=(s_1, a_1, \cdots, a_{n-1}, s_n)$. We call g a *generalized reward function*. In our paper we wish to maximize the expected generalized reward over the infinite future.

This paper clarifies the essence of stochastic dynamic programming problems, and especially exposes the important properties peculiar to a broad class of reward functions which appear in those problems. Such the properties will be introduced, in Section 5, as a recursiveness, a monotonicity and a Lipschitz condition. The recursiveness together with the monotonicity is a basement of Bellman's "principle of optimality" and moreover the Lipschitz condition guarantees the existence of an optimal policy and of the solution to an optimality equation. In Section 5, six examples of generalized reward functions having the recursive property are given, while actually we can conceive infinitely many of them.

* Department of Statistics, University of California, Berkeley and Kyushu University, Fukuoka.

** Department of Mathematics, Kyushu University, Fukuoka.

Blackwell [1] formulated the dynamic programming problem of maximizing the expected, discounted sum over the infinite future just in the form of Markovian decision processes, and gave many elegant results. After that, Strauch [2] treated two problems of maximizing the expected, discounted sum and of maximizing the expected, undiscounted but negative sum, in a parallel way. Performance indices, which they wished to maximize in their papers, have the recursive property in our sense (cf. Section 5). The conception of the monotonicity which we shall impose on the generalized reward function is a generalization of the " monotonicity of operators " in their sense. The Lipschitz condition with coefficients whose infinite product converges to zero will play the role similar to the one of "discounting" in their papers. Many of our proofs are owing to their papers, especially to Balckwell [1].

In Section 2, we prepare the probabilistic definitions and notation used throughout the paper. Section 3 is devoted to formulate our decision problems associated with the generalized reward functions. For every $\varepsilon > 0$, there always exists a (p, ε)-optimal policy (Section 4). In Section 5, we shall define a recursiveness, a monotonicity and a Lipschitz condition, and we shall give an existence theorem for a (p, ε)-optimal Markov policy, assuming the recursiveness and Lipschitz condition to the generalized reward function. Examples of recursive reward functions, too, will be given in Section 5. In Section 6 we shall define an operator associated with measurable mapping f from S into A and define an operator associated with a Markov policy. Then it will be shown that there exists a (p, ε)-optimal stationary policy under two additional assumptions (the monotonicity and stationarity of g). The relation of the optimal return to the optimality equation will be derived along the same line as in Blackwell [1]. Moreover we shall give the theorem that if the action space A is countable, then for any $\varepsilon > 0$ there exists an ε-optimal stationary policy, and that if A is finite, then there exists an optimal stationary policy.

2. Probabilistic definitions and notation.

In this section we shall give the basic notation and definitions to be used throughout the paper. These follow to those of [1].

By a *Borel set* we mean a Borel subset of some Polish space. A polish space is the complete separable metric space. Unless otherwise noted, measurable means with respect to the σ-field of Borel subsets of X, which is expressed by $\mathscr{B}(X)$. A *probability* on a non-empty Borel set X is a probability measure defined over the measurable space $(X, \mathscr{B}(X))$; the set of all probabilities on X is denoted by $P(X)$. For any non-empty Borel sets X, Y a *conditional probability* on Y given X is a function $q(\cdot|\cdot)$ such that for each $x \in X$, $q(\cdot|x)$ is a probability on Y and for each $B \in \mathscr{B}(Y)$, $q(B|\cdot)$ is a Baire function on X. The set of all conditional probabilities on Y given X is denoted by $Q(Y|X)$. The product space of X and Y is denoted by XY. The set of bounded Baire functions on X is denoted by $M(X)$. If $u, v \in M(X)$, $u \geqq v$ means $u(x) \geqq v(x)$ for all $x \in X$. For any $q \in Q(Y|X)$ and q-integrable Baire function on XY, qu denotes the Baire function on X whose value at $x \in X$ is $qu(x)$

$=\int_Y u(x, y)dq(y|x)$. For any $p\in P(X)$ and any $u\in M(X)$, pu is the integral of u with respect to p. For any $p\in P(X)$, $q\in Q(Y|X)$, pq is the probability on XY such that, for every $u\in M(XY)$, $pq(u)=p(qu)$. Every probability m on XY has a factorization $m=pq$; p is unique and is just the marginal distribution of the first coordinate variable with respect to m; q is not unique; it is a version of the conditional distribution of the second coordinate variable given the first.

We can extend the above notation in an obvious way to a finite or countable sequence of non-empty Borel sets. The details are ommitted. A $p\in P(X)$ is degenerate if it is concentrated at some one point $x\in X$; a $q\in Q(Y|X)$ is degenerate if each $q(\cdot|x)$ is so. The degenerate q are exactly those for which there is a Baire function f mapping X into Y for which $q(f(x)|x)=1$ for all $x\in X$. Any such f will also denote its associated degenerate q, so that, for any $u\in M(XY)$, $fu(x)=u(x, f(x))$ for all $x\in X$.

We shall use the following.

LEMMA 2.1 (Blackwell [1]). *For any* $q\in Q(Y|X)$, $\varepsilon>0$ and q-integrable Baire function u on XY, *there is a degenerate* $f\in Q(Y|X)$ *such that*

$$fu\geqq qu \qquad \text{for all} \quad x\in X$$

and

$$q(\{y\,;\ u(x, y)\geqq u(x, f(x))+\varepsilon\}\,|x)=0 \qquad \text{for all} \quad x\in X\,.$$

3. Decision problem definitions.

A Markovian decision problem is defined by S, A, q, g, where S, A are non-empty Borel sets, $q\in Q(S|SA)$, and g a Baire function on $[\bigcup_{n=2}^{\infty}\bigcup_{1\leqq m<n}[S_m\times A_m\times S_{m+1}\times A_{m+1}\times\cdots\times S_n\times R]\cup\bigcup_{m=1}^{\infty}[S_m\times A_m\times S_{m+1}\times A_{m+1}\cdots]]$ where $S_n=S$, $A_n=A$ for $n=1, 2, \cdots$, and $R=(-\infty, \infty)$. A *policy* π is a sequence $\{\pi_1, \pi_2, \cdots\}$, where $\pi_n\in Q(A|H_n)$ and $H_n=SA\cdots S$ ($2n-1$ factors) is the set of possible histories of the system when the n-th actmust be chosen. A policy π is (non-randomized) *Markov* if each π_n is a degenerate element of $Q(A|S)$, i. e. $\pi=\{f_1, f_2, \cdots\}$, where each f_n is a measurable function from S into A, and is (non-randomized) *stationary* if there is a measurable function f mapping S into A such that $\pi_n=f$ for all n. The stationary policy defined by f is denoted by $f^{(\infty)}$. For any policies π and σ, let $\pi^n\sigma=\{\pi_1, \pi_2, \cdots, \pi_n, \sigma_{n+1}, \sigma_{n+2}, \cdots\}$ denote the policy which follows π for the first n stages then switches to σ. For any policy $\pi=\{\pi_1, \pi_2, \cdots\}$, ${}^n\pi$ denotes $\{\pi_{n+1}, \pi_{n+2}, \cdots\}$. In particular ${}^0\pi=\pi$. We shall denote the sequence of policies by $\{\pi^n\}$.

Any policy π, together with the law of motion q of the system, defines for each initial state s a conditional probability on the set $ASAS\cdots$ of futures of the system, i. e. it defines an element of $Q(ASAS\cdots|S)$, namely $e_\pi=\pi_1 q\pi_2 q\cdots$. E^π and $E^{\{\pi_m, \pi_{m+1}, \ldots, \pi_n\}}$ denote the integral operators by e_π and $\pi_m q\pi_{m+1}q\cdots\pi_n q$ respectively. Denote the coordinate functions on $SASA\cdots$ by $s_1, a_1, s_2, a_2, \cdots$ so that our reward on the n-th day, as a function of the past history up to date of the system, is $g(s_1, a_1, \cdots, a_{n-1}, s_n, v(s_1, a_1, \cdots, a_{n-1}, s_n))$, and our reward over the infinite future, as a function on the history $H=SASAS\cdots$, is $g(s_1, a_1, s_2, a_2, \cdots)$. An expected reward function on S from policy π then is given by

$$I(\pi)=e_\pi g=e_\pi[g(s_1, a_1, s_2, a_2, \cdots)] .$$

For any $v\in M(H_n)$, we shall denote by $I_n(\pi, v)$ the expected reward we could expect if we terminate at the n-th stage to receive a generalized terminal reward $v(h_n)=v(s_1, a_1, \cdots, s_n)$ at the terminal state. Thus

$$I_n(\pi, v)=e_\pi g(s_1, a_1, \cdots, s_n, v(s_1, a_1, \cdots, s_n))=\pi_1 q\pi_2 q \cdots \pi_{n-1}qg .$$

We shall denote $I_n(\pi, 0)$ by $I_n(\pi)$. Here it should be noted that $u\leqq v$ may not imply $I_n(\pi, u)\leqq I_n(\pi, v)$ for all π and all n.

For any $p\in P(S)$, and any $\varepsilon>0$, π^* will be called *(p, ε)-optimal* if $p\{I(\pi^*)\geqq I(\pi)-\varepsilon\}=1$ for every π. π^* will be called *ε-optimal* if it is (p, ε)-optimal for every $p\in P(S)$, or equivalently, if $I(\pi^*)\geqq I(\pi)-\varepsilon$ for all π, and will be called *optimal* if it is ε-optimal for every $\varepsilon>0$, or equivalently, if $I(\pi^*)\geqq I(\pi)$ for all π.

Throughout this paper we assume $E^{n\pi}g(h_{n+1})$ is uniformly bounded for all π, n and h_{n+1}.

4. Existence of a (p, ε)-optimal policy.

The proof of the following theorem is similar to that of Theorem 1 in [1].

THEOREM 4.1. *For any $p\in P(S)$ and any $\varepsilon>0$ there is a (p, ε)-optimal policy.*

Let $v^*=\sup_\pi I(\pi)$. Then, in general, v^* may not be measurable. We call the policy π^* *strongly (p, ε)-optimal* if it holds that

$$p\{I(\pi^*)\geqq v^*-\varepsilon\}=1 . \tag{4.1}$$

Obviously, the strong (p, ε)-optimality implies the (p, ε)-optimality. Note that the set $\{s\,;\, I(\pi^*)\geqq v^*-\varepsilon\}$ is in general not Borel. Strauch [2], however, has shown that it was Borel in the completion of the Borel sets with respect to every $p\in P(S)$. Hence the statement (4.1) has the meaning. Furthermore, his method of proof enables us to show

THEOREM 4.2. *For any $p\in P(S)$ and any $\varepsilon>0$ there exists a strongly (p, ε)-optimal policy.*

5. Recursiveness, monotonicity and Lipschitz condition.

Throughout the remainder of this paper, we assume that g satisfies the *recursive relations*; for any π and $1\leqq m\leqq n$

$$E^{\{\pi_m, \pi_{m+1}, \cdots, \pi_n, \cdots\}}g=E^{\{\pi_m, \cdots, \pi_n\}}g(s_m, a_m, \cdots, s_{n+1}, E^{n\pi}g) . \tag{5.1}$$

These relations mean that for $1\leqq m\leqq n$,

$$\pi_m q\pi_{m+1}q \cdots (h_m)g({}^{m-1}h)$$
$$=\pi_m q\pi_{m+1}q \cdots \pi_n q(h_m)g(s_m, a_m, \cdots, s_{n+1}, \pi_{n+1}q\pi_{n+2}q \cdots g({}^n h)) ,$$

where ${}^k h=(s_{k+1}, a_{k+1}, s_{k+2}, \cdots)$ and $h_k=(s_1, a_1, \cdots, s_k)$ for $k=1, 2, \cdots$, and ${}^0 h=h$. The generalized reward function g is called the *recursive reward function* if g satisfies recursive relations. Note that putting $m=1$ in (5.1) yields

$$I(\pi)=I_n(\pi, I(^n\pi)).$$

Before proceeding to (p, ε)-domination we shall give six examples of g which is the recursive reward function. The reward functions in the following examples are those which we usually encounter in the dynamic programming problems. (cf. Bellman [4]). The recursiveness in these examples occures from the peculiarity of the type of g: the accumulation (in some wide sense) of stage-wise rewards.

EXAMPLE 1. (Additive process)

$$\begin{aligned} &g(s_m, a_m, \cdots, s_n, v(s_1, a_1, \cdots, s_n))=\sum_{k=m}^{n-1}\beta^{k-m}r_k(s_k, a_k, s_{k+1})+\beta^{n-m}v(s_1, a_1, \cdots, s_n) \\ &g(s_m, a_m, s_{m+1}, a_{m+1}, \cdots)=\sum_{k=m}^{\infty}\beta^{k-m}r_k(s_k, a_k, s_{k+1}) \end{aligned} \tag{5.2}$$

where $0\leqq\beta\leqq1$, $r_k\in M(SAS)$, $v\in M(H_n)$ for $k=1, 2, \cdots$. If $0\leqq\beta<1$ and if for some $K<\infty$, $\|r_k\|\leqq K$ $(k=1, 2, \cdots)$, then g in (5.2) is well-defined. The case when $0\leqq\beta<1$ and $r_k=r$ for $k=1, 2, \cdots$, consequently g becomes to be well-defined, has been called the discounted case.

EXAMPLE 2. (Absorbed process)

We consider the additive process $(S, A, q, \{r_k\}, v, \beta=1)$ where there exist a state $s_0\in S$ and a positive $\alpha<1$ such that for any action $a\in A$ and any state $s\in S$

$$q(\{s_0\}\mid s, a)\geqq\alpha.$$

It is convienient to call this process the absorbed process. Let's define new transition law q' as follows:

$$q'(B\mid s, a)=\begin{cases} \dfrac{q(B\mid s, a)}{1-\alpha} & \text{for} \quad s_0\notin B \\ \dfrac{q(B\mid s, a)-\alpha}{1-\alpha} & \text{for} \quad s_0\in B. \end{cases}$$

Note that q' is a Markov transition law. It is easy to show that the absorbed process with $\{r_k\}$ satisfying $r_k(s, a, s_0)=0$ for all $k\geqq1$ and all $(s, a)\in SA$ can be reduced to the additive process $(S, A, q', \{r_k\}, v, \beta')$ with the discount factor $\beta'=1-\alpha<1$. This is essentially the result of the fact

$$\int_S r_k(s_k, a_k, s_{k+1})dq(s_{k+1}\mid s_k, a_k)=(1-\alpha)\int_S r_k(s_k, a_k, s_{k+1})dq'(s_{k+1}\mid s_k, a_k).$$

Furthermore, if $r_k=r$ for $k=1, 2, \cdots$, then all results established for the discounted Markovian decision processes remain valid for the absorved process, because the absorved process $(S, A, q, \{r_k\}, v)$ is equivalent to new discounted process $(S, A, q', \{r_k\}, v, \beta')$.

EXAMPLE 3. (Multiplicative process)

$$g(s_m, a_m, \cdots, s_n, v(s_1, a_1, \cdots, s_n))=\prod_{k=m}^{n-1}r_k(s_k, a_k, s_{k+1})v(s_1, a_1, s_2, a_2, \cdots, s_n),$$

$$g(s_m, a_m, s_{m+1}, a_{m+1}, \cdots)=\prod_{k=m}^{\infty}r_k(s_k, a_k, s_{k+1}),$$

where

$$v \in M(H_n) \quad \text{and} \quad r_k \in M(SAS) \qquad \text{for} \quad k=1, 2, \cdots .$$

We assume herein that $e_\pi g$ is well-defined for every π. We sometimes encounter the multiplicative process in a field of system reliability problems. For example, suppose that the system consists of a series of state, where an arbitrary large number of duplicate components are used. In this system $r_k(s, a, s') = 1-(1-p_k^{n_k(s,a)})$ is the probability of successful performance at the k-th stage, provided that taking an action a at state s and the k-th stage implies we use $n_k(s, a)$ duplicate components. Here p_k is the probability of success in each component at k-th stage. Then, $I(\pi)$ is the overall probability of successful operation of the system, when a policy π is used.

EXAMPLE 4. (Multiplicated additive process)

$$g(s_m, a_m, \cdots, s_n, v(s_1, a_1, \cdots, s_n))$$
$$= \sum_{j=m}^{n-1} \prod_{k=m}^{j} r_k(s_k, a_k, s_{k+1}) + \prod_{k=m}^{n-1} r_k(s_k, a_k, s_{k+1}) v(s_1, a_1, \cdots, s_n),$$

$$g(s_m, a_m, s_{m+1}, a_{m+1}, \cdots) = \sum_{j=m}^{\infty} \prod_{k=m}^{j} r_k(s_k, a_k, s_{k+1}).$$

EXAMPLE 5.

$$g(s_m, a_m, \cdots, s_n, v(s_1, a_1, \cdots, s_n))$$
$$= \sum_{j=m}^{n-1} \{[1-r_j(s_j, a_j, s_{j+1})] e^{\sum_{k=m}^{j} r_k(s_k, a_k, s_{k+1})}\} + v(s_1, a_1, \cdots, s_n) e^{\sum_{k=m}^{n-1} r_k(s_k, a_k, s_{k+1})},$$

$$g(s_m, a_m, s_{m+1}, a_{m+1}, \cdots) = \sum_{k=m}^{\infty} \{[1-r_j(s_j, a_j, s_{j+1})] e^{\sum_{k=m}^{j} r_k(s_k, a_k, s_{k+1})}\},$$

where

$$v \in M(H_n) \quad \text{and} \quad r_n \in M(SAS) \qquad \text{for} \quad n=1, 2, \cdots .$$

EXAMPLE 6.

$$g(s_m, a_m, s_{m+1}, v(s_1, a_1, \cdots, s_{m+1})) = h(r_m(s_m, a_m, s_{m+1})) + \frac{v(s_1, a_1, \cdots, s_{m+1})}{r_m(s_m, a_m, s_{m+1})},$$

$$g(s_m, a_m, \cdots, s_n, v(v_1, a_1, \cdots, s_n))$$
$$= h(r_m(s_m, a_m, s_{m+1})) + \sum_{k=m}^{n-2} \frac{h(r_{k+1}(s_{k+1}, a_{k+1}, s_{k+2}))}{\prod_{j=m}^{k} r_j(s_j, a_j, s_{j+1})} + \frac{v(s_1, a_1, \cdots, s_n)}{\prod_{j=m}^{n-1} r_j(s_j, a_j, s_{j+1})}$$

$$\text{for} \quad m+2 \leqq n,$$

$$g(s_m, a_m, s_{m+1}, a_{m+1}, \cdots) = h(r_m(s_m, a_m, s_{m+1})) + \sum_{k=m}^{\infty} \frac{h(r_{k+1}(s_{k+1}, a_{k+1}, s_{k+2}))}{\prod_{j=m}^{k} r_j(s_j, a_j, s_{j+1})},$$

where

$$v \in M(H_n), \quad h \in M(R) \quad \text{and} \quad r_k \in M(SAS) \qquad \text{for} \quad k=1, 2, \cdots .$$

It is easy to show that these examples satisfy the recursive relations.

We now define a monotonicity and a Lipschitz condition for g.

DEFINITION 5.1. The generalized reward function g is *monotone* if $u \leqq v$ $(u, v \in M(S))$ implies that for any π, all h_n and all $n \geqq 1$

$$\pi_n q(h_n) g(s_n, a_n, s_{n+1}, u(s_{n+1})) \leqq \pi_n q(h_n) g(s_n, a_n, s_{n+1}, v(s_{n+1})) .$$

DEFINITION 5.2. The generalized reward function g satisfies a *Lipschitz condition* if there exists a sequnce $\{K_n\}$ of positive numbers such that

$$\begin{aligned} &\|\pi_n qg(s_n, a_n, s_{n+1}, u(s_1, a_1, \cdots, s_{n+1})) \\ &\qquad\qquad -\pi_n qg(s_n, a_n, s_{n+1}, v(s_1, a_1, \cdots, s_{n+1}))\| \\ &\quad \leqq K_n \|u-v\| \end{aligned}$$

for all $u, v \in M(H_{n+1})$, all $\pi_n \in Q(A|H_n)$ and every $n \geqq 1$.

Note that the norm $\|\cdot\|$ at the left hand side in above definition is

$$\begin{aligned} &\sup_{h_n \in H_n} |\pi_n qg(s_n, a_n, s_{n+1}, u(s_1, a_1, \cdots, s_{n+1}))(h_n) \\ &\qquad\qquad -\pi_n qg(s_n, a_n, s_{n+1}, v(s_1, a_1, \cdots, s_{n+1}))(h_n)| . \end{aligned}$$

Now we put the following assumptions.

ASSUMPTION (I). g is monotone.

ASSUMPTION (II). g satisfies the Lipschitz condition with the Lipschitz coefficients such that $K_1 K_2 \cdots K_N \to 0$ as $N \to \infty$.

THEOREM 5.1. *Under Assumption* (II), *for any* $p \in P(S)$, $\varepsilon > 0$ *and any* π *there exists a Markov policy* π^* *which* (p, ε)*-dominates* π, *i.e.* $p\{I(\pi^*) \geqq I(\pi) - \varepsilon\} = 1$.

PROOF. If any two policies π, π' agree for the first N stages, i.e., if $\pi = \{\pi_1, \pi_2, \cdots, \pi_N, \pi_{N+1}, \pi_{N+2}, \cdots\}$ and $\pi' = \{\pi_1, \pi_2, \cdots, \pi_N, \pi'_{N+1}, \pi'_{N+2}, \cdots\}$ then

$$\begin{aligned} \|I(\pi) - I(\pi')\| &= \|\pi_1 qg(s_1, a_1, s_2, E^{1\pi}g) - \pi_1 qg(s_1, a_1, s_2, E^{1\pi'}g)\| \\ &\leqq K_1 \|E^{1\pi}g - E^{1\pi'}g\| \\ &= K_1 \|\pi_2 qg(s_2, a_2, s_3, E^{2\pi}g) - \pi_2 qg(s_2, a_2, s_3, E^{2\pi'}g)\| \\ &\leqq K_1 K_2 \|E^{2\pi}g - E^{2\pi'}g\| \\ &\quad \vdots \\ &\leqq K_1 K_2 \cdots K_N \|E^{N\pi}g - E^{N\pi'}g\| \leqq K_1 K_2 \cdots K_N (\|E^{N\pi}g\| + \|E^{N\pi'}g\|) \\ &\leqq 2K_1 K_2 \cdots K_N \cdot L . \end{aligned}$$

where L is a finite number depending on none of π, π' and N. Hence we may suppose that π is already Markov from some stage on, say for $n > N$, namely $\pi = \{\pi_1, \pi_2, \cdots, \pi_N, f_{N+1}, f_{N+2}, \cdots\}$. By the recursiveness of g,

$$E^{\pi}g = E^{(\pi_1, \cdots, \pi_{N-1})}g(s_1, a_1, \cdots, s_N, E^{N-1\pi}g) .$$

On the other hand, by Lemma 2.1, for $\gamma = \frac{\varepsilon}{N}$ there exists an f_N such that

$$\begin{aligned} E^{N-1\pi}g &= \pi_N qg(s_N, a_N, s_{N+1}, E^{N\pi}g(s_{N+1})) \\ &\leqq f_N qg(s_N, a_N, s_{N+1}, E^{N\pi}g(s_{N+1})) + \gamma \\ &\qquad\qquad \text{with } p\pi_1 q \cdots \pi_{N-1}q\text{-prob. } 1 . \end{aligned}$$

That is,

$$\pi_{N-1}q\pi_N qg(s_N, a_N, s_{N+1}, E^{N\pi}g) \leqq \pi_{N-1}q[f_N qg(s_N, a_N, s_{N+1}, E^{N\pi}g)+\gamma],$$

or equivalently

$$\pi_{N-1}qg(s_{N-1}, a_{N-1}, s_N, E^{N-1\pi}g) \leqq \pi_{N-1}qf_N qg(s_N, a_N, s_{N+1}, E^{N\pi}g)+\gamma$$

with $p\pi_1 q \cdots \pi_{N-2}q$-prob. 1.

Using this procedure N times will produce a Markov $\pi^* = \{f_1, f_2, \cdots\}$ such that

$$E^{\pi}g \leqq E^{(f_1, f_2, \cdots, f_N, {}^N\pi)}g + N\gamma = E^{\pi^*}g + \varepsilon \qquad \text{with } p\text{-prob. } 1.$$

This completes the proof.

COROLLARY. *Under Assumption* (II), *for any* $\varepsilon > 0$ *there exists a* (p, ε)*-optimal Markov policy* π^*.

6. Stationary policies and operators.

We shall say that g is *stationary* if for every $m \geqq 1$ and $k \geqq 1$

$$g | S_m \times A_m \times S_{m+1} \times R = g | S_{m+k} \times A_{m+k} \times S_{m+k+1} \times R$$

by identifying $S_m \times A_m \times S_{m+1} \times R$ and $S_{m+k} \times A_{m+k} \times S_{m+k+1} \times R$, where $g|E$ denotes the restriction of g on E.

Throughout this section we assume that g is stationary. The stationarity of g yields that $r_1 = r_2 = \cdots = r_n = \cdots =$ (say r) in every example stated in Section 5.

DEFINITION 6.1. For each measurable function f mapping S into A, and $u \in M(S)$, let

$$\begin{aligned} T_f u(s) &= fqg(s, a, s', u(s')) \\ &= \int_S g(s, f(s), s', u(s')) dq(s' | s, f(s)). \end{aligned}$$

It is easily verified that the operator T_f associated with measurable function f mapping S into A is a mapping from $M(S)$ into $M(S)$.

Now we shall express the operator T_f explicitly for each example. For Example 1, also for Example 2, we have

$$\begin{aligned} T_f u(s) &= fq[r + \beta u](s) \\ &= \int_S [r(s, f(s), s') + \beta u(s')] dq(s' | s, f(s)). \end{aligned}$$

The operator T_f in this case was originally given by Blackwell [1].

For Example 3, we have

$$\begin{aligned} T_f u(s) &= fq[r \cdot u](s) \\ &= \int_S [r(s, f(s), s') u(s')] dq(s' | s, f(s)). \end{aligned}$$

For Example 4, we have

$$\begin{aligned} T_f u(s) &= fq[r+r\cdot u](s) \\ &= \int_S [r(s, f(s), s')+r(s, f(s), s')u(s')]dq(s' \mid s, f(s)) . \end{aligned}$$

For Example 5, we have

$$\begin{aligned} T_f u(s) &= fq[(1-r)e^r+e^r u](s) \\ &= \int_S [(1-r(s, f(s), s'))e^{r(s,f(s),s')} \\ &\qquad + e^{r(s,f(s),s')}u(s')]dq(s' \mid s, f(s)) . \end{aligned}$$

And for Example 6 we have

$$\begin{aligned} T_f u(s) &= fq\Big[h(r)+\frac{u}{r}\Big](s) \\ &= \int_S \Big[h(r(s, f(s), s'))+\frac{u(s')}{r(s, f(s), s')}\Big]dq(s' \mid s, f(s)) . \end{aligned}$$

The following two lemmas are immediate consequences of the assumptions.

LEMMA 6.1. *Under Assumption* (I), *the operator T_f is monotone for any measurable f from S into A. That is, $u \leqq v$ implies*

$$T_f u \leqq T_f v .$$

LEMMA 6.2. *Under Assumption* (II), *T_f is a contraction mapping on $M(S)$ with some contraction coefficient $0<K<1$ for all measurable f from S into A.*

We now return to Examples and look over whether the operator T_f is a monotone or contraction mapping.

In Example 1, T_f is always a monotone and contraction mapping on $M(S)$ for $0 \leqq \beta < 1$. If $r \geqq 0$ then, T_f in both of Example 3 and Example 4 is monotone. T_f is always monotone for Example 5. If $\|r\|<1$, then T_f in both of Example 3 and Example 4 is a contraction mapping. In Example 5, if $\|e^r\|<1$ then T_f is a contraction. In Example 6, if $r>0$ then T_f is monotone, and if $\inf_{(s,a,s')\in SAS} |r(s, a, s')|>1$ then T_f is a contraction.

THEOREM 6.1. (a) *Under Assumption* (II) *for any $u \in M(S)$ and constant c, $\|T_f(u+c)-T_f u\| \leqq Kc$.*

(b) *For any Markov policy $\pi=\{f_1, f_2, \cdots\}$, $TI(\pi)=I(f, \pi)$, where $\{f, \pi\}$ denotes the Markov one $\{f, f_1, f_2, \cdots\}$.*

For any Markov $\pi=\{f_1, f_2, \cdots\}$ we shall say that f mapping S into A is *π-generated* if there exists a partition of S into Borel sets $S_1, S_2, \cdots$, such that $f=f_n$ on S_n; we say that a Markov $\pi'=\{g_1, g_2, \cdots\}$ is *π-generated* if each g_n is so. We associate with each Markov $\pi=\{f_1, f_2, \cdots\}$ the operator U_π, mapping $M(S)$ into $M(S)$, defined by $U_\pi u=\sup_{n\geqq 1} T_{f_n}u$.

THEOREM 6.2. (a) *If T_{f_n} is monotone for $n \geqq 1$, then U_π is so.*

(b) *If $\|T_{f_n}u-T_{f_n}v\| \leqq K\|u-v\|$ for $k=1, 2, \cdots$, then $\|U_\pi u-U_\pi v\| \leqq K\|u-v\|$.*

(c) *For any T_f assoicated with π-generated f, $T_f u \leqq U_\pi u$.*

(d) *For any $u \in M(S)$ and any $\varepsilon > 0$, there exists a π-generated f whose associated T_f satisfies $T_f u \geqq U_\pi u - \varepsilon$.*

PROOF. (a) is trivial. (c) is immediate from the definition. For (b), consider arbitrary u, v and s, and write $(U_\pi u)(s) = (U_\pi v)(s) + k$. Consider the case $k > 0$. For any integer n, there exists an m such that

$$T_{f_m} u(s) \geqq U_\pi u(s) - \frac{k}{n}\,. \tag{6.1}$$

On the other hand,

$$U_\pi u(s) - \frac{k}{n} \geqq U_\pi u(s) - k = U_\pi v(s) \geqq T_{f_m} v(s)\,. \tag{6.2}$$

Hence, (6.1), (6.2) and the assumption yield

$$\begin{aligned} 0 \leqq U_\pi u(s) - U_\pi v(s) - \frac{k}{n} &\leqq T_{f_m} u(s) - T_{f_m} v(s) \\ &\leqq K\|u - v\|\,. \end{aligned} \tag{6.3}$$

Since above (6.3) is valid for each n, it follows that

$$|U_\pi u(s) - U_\pi v(s)| \leqq K\|u - v\|\,.$$

This inequality is trivial for $k = 0$ and is similarly established for $k < 0$. Then

$$\|U_\pi u - U_\pi v\| \leqq K\|u - v\|\,.$$

For (d), the proof is similar to Theorem 4 (d) in [1].

THEOREM 6.3. (a) *For any $u \in M(S)$ and $\varepsilon > 0$, there exists a π-generated f whose associated T_f satisfies $\|T_f u - U_\pi u\| \leqq \varepsilon$.*

(b) *For any $u \in M(S)$, $n \geqq 1$ and $\varepsilon > 0$, there exists a π-generated π' such that $\|I_n(\pi', u) - U_\pi^n u\| \leqq \varepsilon$.*

PROOF. (a) is immediate from Theorem 6.2 (d). For (b), then there exists a π-generated f_n such that

$$\|T_{f'_n} u - U_\pi u\| \leqq \varepsilon_n\,.$$

Similarly for $\varepsilon_{n-1} > 0$ there exists π-generated f'_{n-1} such that

$$\|T_{f'_{n-1}}(T_{f'_n} u) - U_\pi(T_{f'_n} u)\| \leqq \varepsilon_{n-1}\,.$$

Then

$$\begin{aligned} \|T_{f'_{n-1}} T_{f'_n} u - U_\pi^2 u\| &\leqq \|T_{f'_{n-1}} T_{f'_n} u - U_\pi(T_{f'_n} u)\| + \|U_\pi(T_{f'_n} u) - U_\pi(U_\pi u)\| \\ &\leqq \varepsilon_{n-1} + K\|T_{f'_n} u - U_\pi u\| \\ &\leqq \varepsilon_{n-1} + K\varepsilon_n\,. \end{aligned}$$

In general there exist $f'_1, f'_2, \cdots, f'_n$ such that for $i = 1, 2, \cdots, n$

$$\|T_{f'_1} \cdots T_{f'_n} u - U_\pi^{n-i+1} u\| \leqq d_i\,,$$

where $d_i = \varepsilon_i + K\varepsilon_{i+1} + \cdots + K^{n-i}\varepsilon_n$. In particular we can choose $\varepsilon_1, \cdots, \varepsilon_n$ so that $d_1 \leqq \varepsilon$. This showes the existence of the desired π'.

Note that Theorem 6.3 (b) implies that for any $u \in M(S)$, $n \geqq 1$ and $\varepsilon > 0$, there exists a π-generated π' such that $I_n(\pi', u) \geqq U_\pi^n u - \varepsilon$.

THEOREM 6.4. *Under Assumptions* (I) *and* (II), U_π *is a contraction mapping on* $M(S)$ *with the contraction coefficient* K, *i.e.* $\|U_\pi u - U_\pi v\| \leqq K\|u-v\|$, *so that, from the Banach fixed-point theorem,* U_π *has a unique fixed point* u^*, *and* $\|U_\pi^n u - u^*\| \leqq K^n \|u-u^*\|$ *for all* n.

PROOF. $v \leqq u + \|u-v\|$ yields

$$U_\pi v \leqq U_\pi(u+\|u-v\|) \leqq U_\pi u + K\|u-v\|$$

using Theorem 6.2 (b). Interchange u and v to obtain

$$U_\pi u \leqq U_\pi v + K\|v-u\| ,$$

completing the proof.

THEOREM 6.5. *Let Assumptions* (I), (II) *be satisfied.*

(a) *For any Markov policy* $\pi = \{f_1, f_2, \cdots\}$, *denoting by* T_{f_n} *associated with* f_n *and by* $U_\pi = \sup_{n \geqq 1} T_{f_n}$, *the operator associated with* π, *the fixed point* u^* *of* U_π *is the optimal return among* π*-generated policies:* $I(\pi') \leqq u^*$ *for every* π*-generated* π', *and for every* $\varepsilon > 0$ *there is a* π*-generated* f *such that* $I(f^{(\infty)}) \geqq u^* - \varepsilon$. *Any* f *with* $T_f u^* \geqq u^* - \varepsilon(1-K)$ *satisfies this inequality. Furthermore, any* f *with* $\|T_f u^* - u^*\| \leqq \varepsilon(1-K)$ *satisfies* $\|I(f^{(\infty)}) - u^*\| \leqq \varepsilon$.

(b) *For any* $p \in P(S)$, $\varepsilon > 0$, *there exists a* (p, ε)*-optimal policy which is stationary.*

(c) *For any* $\varepsilon \geqq 0$, *if there is an* ε*-optimal* $\pi^* = \{\pi_1, \pi_2, \cdots\}$, *there exists an* $\varepsilon/(1-K)$*-optimal policy which is stationary.*

(d) *Denote for each* $a \in A$ *by* T_a *the operator associated with* $f \equiv a$. *Any* u *with* $T_a u \leqq u$ *for all* a *is an upper bounded on reward:* $I(\pi) \leqq u$ *for all* π.

(e) *If for every* $\varepsilon > 0$ *there exists an* ε*-optimal policy, then the optimal return* u^* *is a Baire function and it satisfies the optimality equation* $u^* = \sup_{a \in A} T_a u^*$.

(f) *A policy* π *is optimal if and only if its return* $I(\pi)$ *satisfies the optimality equation.*

PROOF. (a) For any π-generated $\pi' = \{g_1, g_2, \cdots\}$, we have $I(\pi') = T_{g_1} T_{g_2} \cdots T_{g_n} u_n$, where $u_n = I(g_{n+1}, g_{n+2}, \cdots)$. Since each T_{g_i} is a contraction operator with contraction coefficient K,

$$\begin{aligned}\|T_{g_1}T_{g_2}\cdots T_{g_n}u_n - T_{g_1}T_{g_2}\cdots T_{g_n}u^*\| &\leqq K^n\|u_n - u^*\| \\ &\leqq K^n(\|g\| + \|u^*\|) .\end{aligned}$$

Thus $T_{g_1}\cdots T_{g_n}u^* \to I(\pi')$ as $n \to \infty$. But $T_{g_1}\cdots T_{g_n}u^* \leqq U_\pi^n u^* = u^*$, so that $I(\pi') \leqq u^*$. From Theorem 6.3 (a), there is a π-generated f for which $T_f u^* \geqq U_\pi u^* - \varepsilon' = u^* - \varepsilon'$, where $\varepsilon' = \varepsilon(1-K)$. We can verify inductively that

$$T_f^n u^* \geqq u^* - \varepsilon'(1+K+\cdots+K^{n-1}) \qquad \text{for all} \quad n \geqq 1 .$$

Since $T_f^n u^* \to I(f^{(\infty)})$, we conclude that

$$I(f^{(\infty)}) \geqq u^* - [\varepsilon'/(1-K)] = u^* - \varepsilon .$$

Last two inequalities are immediate from Theorem 6.3 (a).

(b) This is immediate from Corollary to Theorem 5.1 and above (a).

(c) For any $\pi^*=\{\pi_1, \pi_2, \cdots\}$, $I(\pi^*)=e_{\pi^*}g=\pi_1 qg(s, a, s_2, E^{1\pi^*}g)$, where $E^{1\pi^*}g=E^{1\pi^*}g(s, a, s_2)=I(\pi_{s,a})(s_2)$ and $\pi_{s,a}$ denotes the policy which π^* specifies starting with the second stage, when the first state and action are s, a respectively, i.e. $\pi_{s,a}=\{\pi'_1, \pi'_2, \cdots\}$ where

$$\pi'_n(\cdot\,|\,s_1, a_1, \cdots, s_n)=\pi_{n+1}(\cdot\,|\,s, a, s_1, a_1, \cdots, s_n)\,.$$

If π^* is ε-optimal, $E^{1\pi^*}g(s, a, s_2)=I(\pi_{s,a})(s_2)\leqq I(\pi^*)(s_2)+\varepsilon$ for all $s_2\in S$, so that, by monotonicity of g

$$\begin{aligned} I(\pi^*)&=\pi_1 qg(s_1, a_1, s_2, E^{1\pi}g)\\ &\leqq \pi_1 qg(s_1, a_1, s_2, I(\pi^*)(s_2)+\varepsilon)\\ &=\pi_1 h\,, \qquad \text{say}\,. \end{aligned}$$

From the Lemma 2.1, there exists an f for which $\pi_1 h\leqq fh$ for all s, namely

$$\begin{aligned} \pi_1 qg(s_1, a_1, s_2, I(\pi^*)(s_2)+\varepsilon)&\leqq fqg(s_1, a_1, s_2, I(\pi^*)(s_2)+\varepsilon)\\ &=T_f(I(\pi^*)+\varepsilon)\,. \end{aligned}$$

Then,

$$I(\pi^*)\leqq T_f(I(\pi^*)+\varepsilon)\leqq T_f I(\pi^*)+K\varepsilon\,.$$

Again,

$$T_f(I(\pi^*))\leqq T_f(T_f(I(\pi^*)+K\varepsilon)\leqq T_f^2 I(\pi^*)+K^2\varepsilon\,.$$

By induction on n we obtain

$$I(\pi^*)\leqq T_f^n I(\pi^*)+K\varepsilon+K^2\varepsilon+\cdots+K^n\varepsilon\,.$$

Letting $n\to\infty$ yields

$$I(\pi^*)\leqq I(f^{(\infty)})+K\varepsilon/(1-K)\,.$$

Since π^* is ε-optimal, $f^{(\infty)}$ is $\varepsilon/(1-K)$-optimal.

(d) The proof of (d) is similar to that of Theorem 6 (d) in [1], and is omitted.

(e) From (c), the hypothesis implies that there exists a $1/n$-optimal stationary policy $f_n^{(\infty)}$, say. With $\pi=\{f_1, f_2, \cdots\}$, the fixed u^* of the U_π, is, from (a), the optimal reward among π-generated policies. In particular $u^*\geqq I(f_n^{(\infty)})$, so that $u^*\geqq I(\pi)$ for all π, and u^* is the optimal return. We have $\sup_{a\in A} T_a u^*\geqq U_\pi u^*=u^*$. On the other hand, for any $a\in A$,

$$\begin{aligned} T_a u^*&\leqq T_a(I(f_n^{(\infty)})+(1/n))\leqq T_a(I(f_n^{(\infty)})+\frac{K}{n}\\ &=I(a, f_n^{(\infty)})+\frac{K}{n}\leqq u^*+\frac{K}{n}\,, \end{aligned}$$

where $(a, f_n^{(\infty)})$ is the Markov policy $\{g, f_n, f_n, \cdots\}$ with $g\equiv a$. Letting $n\to\infty$ yields $T_a u^*\leqq u^*$. Thus u^* satisfies the optimality equation.

(f) The proof is same as in Theorem 6 (f) of [1].

Finally we give a theorem in the case when A is countable or finite, without stating the proof.

THEOREM 6.6. *Let Assumptions* (I) *and* (II) *be satisfied.* (i) *If A is countable, then for any $\varepsilon>0$ there exists an ε-optimal policy which is stationary.* (ii) *If A is finite, then there exists an optimal policy which is stationary.*

References

[1] D. Blackwell, *Discounted dynamic programming,* Ann. Math. Statist., **36** (1965), 226-235.

[2] R. E. Strauch, *Negative dynamic programming,* Ann. Math. Statist., **37** (1966), 871-890.

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

[4] R. Bellman, *Dynamic Programming,* Princeton Univ. Press. (1957).