5

SEMI-MARKOV DECISION PROCESSES WITH COUNTABLE STATE SPACE ANE COMPACT ACTION SPACE

By

Masami YASUDA

(Received October 5, 1977)

Reprinted from the Bulletin of Mathematical Statistics, Research Association of Statistical Sciences, Vol. 18, No. 1~2

FUKUOKA, JAPAN
1978

SEMI-MARKOV DECISION PROCESSES WITH COUNTABLE STATE SPACE ANE COMPACT ACTION SPACE

By

Masami YASUDA*

(Received October 5, 1977)

0. Abstract

We shall be concerned with the optimization problem of semi-Markov decision processes with countable state space and compact action space. Defined is the generalized reward function associated with the semi-Markov decision processes which include the ordinary discounted Markov decision processes of discrete time parameter and also the continuous time Markov decision processes. Main results are (a) the existence of an optimal stationary policy and (b) the relation between the maximal expected reward and the optimality equation. Also (c) some properties of the optimal staionary policy and the principle of optimality are obtained.

1. Introduction and summary of results

Semi-Markov decision processes with countable state space S and compact action space A are considered. A policy π is defined as a sequence of mappings f_n from S^n into $A(n\geqq1)$. For each policy π, $X^\pi(t)$ denotes a state of the system generated by the policy and $A^\pi(t)$ denotes a stochastic process which signifies the utilizing mapping of the policy at time t. These stochastic processes are constructed exactly in section 2. In section 3, 4 and 5, we study the problem of maximizing

$$I(\pi):=E\left[\int_0^{\zeta(\pi)} r(X^\pi(t), A^\pi(t))G(dt)\right]$$

with respect to π, where E denotes an expection operator and, $\zeta(\pi)$ is a killing time in the process, r a given function on SA and G a measure on $[0, \infty)$. The maximal expected reward I^* means $\sup I(\pi)$ where the supremum is taken over all policies. An optimal and an ε-optimal policies are defined. We show in section 4 that both of the families of optimal and of ε-optimal policies can be reduced those of Markov or stationary ones.

* Statistical Laboratory, College of General Education, Chiba University, Chiba 280, Japan.

Under some conditions we get the following results:

(a) there exists an optimal stationary policy f^∞, that is, $I^*=I(f^\infty)$;

(b) the optimal reward which is a function of the initial state satisfies a non-linear functional equation called the optimality equation, and conversely the solution of the optimality equation is the optimal reward;

(c) the optimal stationary policy maximizes, in the family of stationary policies, a conditional reward

$$E\left[\int_s^{\zeta(\pi)} r(X^\pi(t), A^\pi(t))G(dt)\mid \mathcal{F}_s^\pi\right]$$

and an expected reward

$$E\left[\int_s^{\zeta(\pi)} r(X^\pi(t), A^\pi(t))G(dt)\right].$$

These results are discussed in section 5 and 6.

A simple type of Markov decision process was introduced by Bellman [1]. Afterward, Howard [11], Blackwell [2], [3], [4], Maitra [17], [18], Strauch [22], Veinott [23], [24], Hinderer [9] studied more general types of Markov decision processes with discrete time parameter extensively. Analogously Markov decision processes with continuous time parameter are developed by Howard [11], [12], de Leve [6], Martinlöf [19], Miller [20], Veinott [24], Kakumanu [24]. Since semi-Markov processes include discrete time Markov processes and continuous time Markov jump processes, if we formulate semi-Markov decision processes, the deductive argument implies the both study of discrete and continuous time Markov decision processes. The possibility is due to the reward structure, particularly the property of a measure G, and it is similar to an additive functional in the potentional theory refers to Blumenthal and Getoor [5]. Howard [12], Miller [20], Ross [21], Lippman [16] considered average reward semi-Markov decision processes but we do not discuss the average case here.

2. Formulation, construction of stochastic processes

In this section we shall develop the construction of stochastic processes $X^\pi(t)$ and $A^\pi(t)$ underlying the optimization problem of a semi-Markov decision process.

First we give notations frequently used in the subsequent sections. A notation $:=$ means a definition distinguished from an equality. Let $\mathcal{B}(X)$ be the Borel field of a topological space X. $P(X)$ denotes the set of all probability measures defined on $\mathcal{B}(X)$. For any X, Y, $P(Y|X)$ is the set of all conditional probability measures on Y given X, whose element q is written by $q(dy|x)$ or $q(x;dy)$. $M(X)$ denotes the set of all bounded Borel measurable functions, where X is a topological space. If $u, v\in M(X)$, $u=v$, $u\geqq v$ means, respectively, $u(x)=v(x)$, $u(x)\geqq v(x)$ for all $x\in X$. For any $p\in P(X)$, $u\in M(X)$, $pu:=p(u):=\int_X u(x)p(dx)$. For any $q\in P(Y|X)$ and any $u\in M(XY)$, $qu\in M(X)$ whose value at $x\in X$ is

$$qu(x):=u(x,q(x)):=\int_Y u(x,y)\,q(dy|x).$$

Obviously the above notations are extended to a finite or countable sequence. Every function $u\in M(X)$ has a norm $\|u\|:=\sup\{|u(x)|\,;\,x\in X\}$. Note that we shall not distinguish between the notation of a distribution function and that of the measure deduced from it, and vice versa.

DEFINITION 2. 1. A *semi-Markov decision process* consists of seven objects (S, A, p_0, p, r, F, G) of the following properties:

(i) the *state space* S is a non-empty countable set with a discrete topology,

(ii) the *action space* A is a compact metric topological space,

(iii) the *initial distribution* $p_0\in P(S)$,

(iv) the *transition law* $p\in P(S|SA)$,

(v) the *reward function* $r\in M(SA)$,

(vi) $F\in P(R|SR)$, where $R:=(-\infty,\infty)$ and

(vii) G is a distribution function with $G(0)=0$.

Moreover we shall require Assumption 1(1) for p, Assumption 2(1) for r, Assumption 1(2) and 2(2) for F and Assumption 2(3) for G.

DEFINITION 2. 2. We define a policy, a Markov policy and a stationary policy.

(i) A *policy* π is a sequence of mappings $(f_n\,;\,n\geqq 1):=(f_1, f_2, \cdots)$ where each component f_n is a mapping of a product space S^n into A for $n\geqq 1$.

(ii) A *Markov policy* $\pi:=(f_n\,;\,n\geqq 1)$ is a policy in which each f_n is a mapping of S into A for $n\geqq 1$.

(iii) A *stationary policy* $\pi:=(f_n\,;\,n\geqq 1)$ is a Markov policy in which each f_n does not depend on n. If $f_n=f$ for all n, we denote the stationary policy by $f^\infty:=(f, f, \cdots)$.

Let Π be the set of all policies and for $\pi=(f_n)\in\Pi$, ${}^n\pi$ or $(n)\pi$ is defined by ${}^n\pi:=(n)\pi:=(f_{n+1}, f_{n+2}, \cdots)$ $(n\geqq 1)$. If p is a transition law and r is a reward function, then we define $p_f\in P(S|S^{n+1})$, $p_a\in P(S|S)$, $r(f)\in M(S^{n+1})$ as follows;

(i) $p_f(x_0, \cdots, x_n\,;\,dx):=p(x_n, f(x_0, \cdots, x_n)\,;\,dx)$ for a mapping f from S^{n+1} into A,

(ii) $p_a(x_0\,;\,dx):=p(x_0, a\,;\,dx)$ for $a\in A$,

(iii) $r(f)(x_0, \cdots, x_n):=r(x_n, f(x_0, \cdots, x_n))$ where $x_0, \cdots, x_n\in S$.

We now give an intuitive description of a process and reward to be constructed in the model (S, A, p_0, p, r, F, G). An object or some amount of our investment starting from a state $x_0\in S$ at time 0 remains there for a holding time τ_1. The distribution of x_0 is p_0 and that of τ_1 is $F(x_0, 0\,;\,.)$. At that time it jumps to a new position x_1 according to our decision which we choose on the basis of an information of the previous state x_0. The decision means the selection of a mapping $f_1\,;\,S\to A$. The transition from x_0 to x_1 occurs according to the probability distribution $p_{f_1}(x_0\,;\,.)$. The object remains at x_1 untill time τ_2 whose distribution is $F(x_1, \tau_1\,;\,.)$ but conditionary independent of τ_1. Then it jumps to x_2 according to our decision, that is, the selection of a mapping f_2; $S^2 \to A$, which we choose on the basis of the first two state x_0, x_1. Generally it jumps to x_{n+1} according to the transition distribution $p_{f_{n+1}}(x_0, \cdots, x_n; \cdot)$, where f_{n+1} is choosen on the basis of the previous sequence $(x_0, x_1, \cdots, x_n)$. It remains at x_{n+1} untill a time τ_{n+2} whose distribution is $F(x_{n+1}, \tau_{n+1}; \cdot)$ but conditionary independent of τ_{n+1}. Each change of state and each decision generate immediate reward $r(f_1)(x_0)$, $r(f_2)(x_0, x_1), \cdots$. Also combining each holding duration of costs $G(\tau_{n+1})-G(\tau_n)$, the total reward is set up by the policy $\pi=(f_n)$, the sequence of each decision f_n. Our purpose is to select a policy $\pi=(f_n)$ so that we can make the expected total reward $I(\pi)$ as high as possible.

This section is devoted to a rigorous construction of such a process and reward.

Let (S, A, p_0, p, r, F, G) be a semi-Markov decision process. Let $N:=\{1, 2, \cdots\}$ and a product space $\Omega_0:=(SR_+)^N$, a product σ-algebra $\mathcal{G}_0:=(\mathcal{B}(S)\mathcal{B}(R_+))^N$ where $R_+:=[0, \infty]$ as usual. Thus $(\Omega_0, \mathcal{G}_0)$ is the usual infinite product measurable space over $(SR_+, \mathcal{B}(S)\mathcal{B}(R_+))$. A point $\omega\in\Omega_0$ is a sequence $\{(x_n, t_n); n\geqq 0\}$. Let $Y_n(\omega):=(x_n, t_n)$, $Z_n(\omega):=x_n$ and $\tau_n(\omega):=t_n$. Thus Y_n is the n-th coordinate map and $Y_n=(Z_n, \tau_n)$. We invoke a theorem of Ionescue Tulcea which states the following in the present situation: for a policy $\pi=(f_n)\in\Pi$, there exists a probability measure P^π on $(\Omega_0, \mathcal{G}_0)$ such that

(a) $P^\pi(Y_{n+1}\in C \mid Y_0, \cdots, Y_n)=\int_C p_{f_{n+1}}(Z_0, \cdots, Z_n; dx)F(Z_n, \tau_n; dt)$
for $C\in\mathcal{B}(S)\mathcal{B}(R_+)$, $n\geqq 0$,

(b) $P^\pi(Z_0\in D)=\int_D p_0(dx)$ for $D\in\mathcal{B}(S)$ and

(c) $P^\pi(\tau_0=0)=1$.

Next we shall consider an infinite product space $(\Omega, \mathcal{G}):=\mathrm{X}_{\pi\in\Pi}(\Omega_0^\pi, \mathcal{G}_0^\pi)$ where $\Omega_0^\pi:=\Omega_0$, $\mathcal{G}_0^\pi:=\mathcal{G}_0$ for all $\pi\in\Pi$. It holds, by the same theorem, that

LEMMA 2. 1. *For a policy $\pi=(f_n)$, there exists a probability measure P on $(\Omega, \mathcal{G})$ and random variables $Y_n^\pi=(Z_n^\pi, \tau_n^\pi)$ such that*

(a) $P(Y_{n+1}^\pi\in C \mid Y_0^\pi, \cdots, Y_n^\pi)=\int_C p_{f_{n+1}}(Z_0^\pi, \cdots, Z_n^\pi; dx)F(Z_n^\pi, \tau_n^\pi; dt)$
for $C\in\mathcal{B}(S)\mathcal{B}(R_+)$, $n\geqq 0$,

(b) $P(Z_0^\pi\in D)=\int_D p_0(dx)$ for $D\in\mathcal{B}(S)$ and

(c) $P(\tau_0^\pi=0)=1$.

Throughout the paper the expectation E means the integral operator by the probability measure P.

The following assumption is needed for Definition 2.3.

ASSUMPTION 1. For any $a\in A$, $x\in S$, $t\in R$,
(1) $p_a(x; \{x\})=0$,
(2) $F(x, t; B)=0$ if $B\subset(-\infty, t]$.

LEMMA 2. 2. *Under Assumption* 1,
(a) $P(Z_{n+1}^\pi=Z_n^\pi)=0$,
(b) $P(\tau_{n+1}^\pi=\tau_n^\pi)=0$ for all n.

PROOF. (a) From Assumption 1(1), it follows that $P(Z_{n+1}^\pi=Z_n^\pi \mid Z_0^\pi, \cdots, Z_n^\pi)=P^\pi(Z_{n+1}=Z_n \mid Z_0, \cdots, Z_n)=p_{f_{n+1}}(Z_0, \cdots, Z_n; \{Z_n\})=0$. Hence $P(Z_{n+1}^\pi=Z_n^\pi)=E[P(Z_{n+1}^\pi=Z_n^\pi \mid Z_0^\pi, \cdots, Z_n^\pi)]=0$. (b) is proved similarly from Assumption 1(2).

Consequently, for each $\pi\in\Pi$, let $\Omega'_\pi:=\{Z_{n+1}^\pi\neq Z_n^\pi$ and $\tau_{n+1}^\pi>\tau_n^\pi$ for all n and $\tau_0^\pi=0\}$, then $\Omega'_\pi\in\mathcal{G}$ and $P(\Omega'_\pi)=1$ under Assumption 1. Neglecting a set of the measure zero, we can assume that, for each $\pi\in\Pi$, Ω_0^π be the set of all sequences $\{(x_n, t_n); n\geqq 0\}$ with $0=t_0<t_1<\cdots$ and $x_{n+1}\neq x_n$ for all $n\geqq 0$. Let $\Omega:=\mathrm{X}_{\pi\in\Pi}\Omega_0^\pi$ and let $\mathcal{G}$ be the σ-algebra in Ω generated by the coordinate mappings $\{Y_n^\pi; n\geqq 0\}$ for each $\pi\in\Pi$. The measure P is regarded as a probability measure on this $(\Omega, \mathcal{G})$. Hence the following is well defined for all $\omega\in\Omega$ under Assumption 1.

DEFINITION 2. 3. Let $\zeta^\pi(\omega):=\lim_n \tau_n(\omega)$ and then define for $t\geqq 0$

$$X_t^\pi(\omega):=\begin{cases} Z_n^\pi(\omega) & \text{if } \tau_n^\pi(\omega)\leqq t<\tau_{n+1}^\pi(\omega), \\ \Delta_S & \text{if } \zeta^\pi(\omega)\leqq t, \end{cases}$$

$$A_t^\pi(\omega):=\begin{cases} f_{n+1}(Z_0^\pi(\omega), \cdots, Z_n^\pi(\omega)) & \text{if } \tau_n^\pi(\omega)\leqq t<\tau_{n+1}^\pi(\omega), \\ \Delta_A & \text{if } \zeta^\pi(\omega)\leqq t \end{cases}$$

where Δ_S, Δ_A is an artificial point added to S, A respectively in the usual convention.

We use the notations $X^\pi(t)$, $A^\pi(t)$ or X_t^π, A_t^π dropping out the variable $\omega\in\Omega$. Random variables Y_n^π, X_t^π generate σ-algebras;

$$\mathcal{G}_n^\pi:=\sigma\{Y_m^\pi; 0\leqq m\leqq n\},$$

$$\mathcal{F}_t^\pi:=\sigma\{X_s^\pi; 0\leqq s\leqq t\}.$$

DEFINITION 2. 4. Let, for a policy $\pi\in\Pi$, a stochastic process $\{R_\pi(s); s\geqq 0\}$ be defined by

$$R_\pi(s):=\int_s^{\zeta(\pi)} r(X^\pi(t), A^\pi(t))G(dt)$$

where $\zeta(\pi):=\zeta^\pi$.

This process means the total reward starting at time s and its expectation $E[R_\pi(s)]$, called an expected total reward, will be considered in the next section.

ASSUMPTION 2. (1) $r(\Delta_S, a)=0$ for all $a\in A\cup\{\Delta_A\}$. (2) There exists a distribution function F_x with a parameter $x\in S$ such that

$$\int_{-\infty}^{\infty} u(s)F(x, t; ds)=\int_0^\infty u(s+t)F_x(ds)$$

for all $t\in R_+$ and $u\in M(R_+)$.

Let Σ be the set of zero and increasing points of $F_{x_1}*\cdots*F_{x_n}$ with $x_m\in S(0\leqq m\leqq n)$ for each $n\geqq 0$, where $*$ is a convolution. A point t is an increasing point of a distribution F iff $F\{I\}>0$ for every open interval I containing t. We designate the set of all increasing points of F by $\mathrm{Inc}\{F\}$. Hence

$$\Sigma := \cup_{n\in N}\cup_{x_1\in S}\cdots\cup_{x_n\in S}\mathrm{Inc}\{F_{x_1}*\cdots*F_{x_n}\}\cup\{0\}.$$

(3) There exists a function γ on R_+ which satisfies;

$$G(t+s)=G(s)+\gamma(s)G(t) \quad s,t\in\Sigma.$$

Note. (a) If $F_x(x\in S)$ are unit distributions concentrated at 1, then the decision process becomes a discrete time parameter case; the above Σ and $G(t)$ are

$$\Sigma=\{0,1,2,\cdots\},$$

$$G(t)=(1-\alpha)\sum_{k=0}^{[t]-1}\alpha^k, \quad t\in R_+$$

where $\alpha:=\gamma(1)<1$.

(b) If F_x $(x\in S)$ are expotential distributions with a parameter $\lambda(x)$, then the decision process becomes a continuous time parameter case; the above Σ and $G(t)$ are

$$\Sigma=R_+$$

$$G(t)=1-e^{-\alpha t}, \quad t\in R_+$$

where $\alpha:=-\log\gamma(1)>0$.

We can derive the following lemmas in a simple manner by using the definitions and assumptions.

LEMMA 2. 3. *For a policy* $\pi=(f_n)$,

(a) $R_\pi(S)=\int_s^\infty r(X^\pi(t),A^\pi(t))G(dt),\quad s\in R_+$,

(b) $R_\pi(\tau_n)=r(Z_n^\pi,f_{n+1}(Z_0^\pi,\cdots,Z_n^\pi))\{G(\tau_{n+1}^\pi)-G(\tau_n^\pi)\}+R_\pi(\tau_{n+1}^\pi)$ *for* $n\geqq 0$.

LEMMA 2. 4. (a) $F_x(G):=\int_0^\infty G(t)F_x(dt)=\int_0^\infty\{1-F_x(t)\}G(dt)$

(b) $\int_0^\infty G(s+t)F_x(dt)-G(s)=\gamma(s)F_x(G)$ *for* $s\in\Sigma$.

(c) γ *is a nonincreasing function with for* $\gamma(0)=1$, $\gamma(s)>0$ *for* $0\leqq s<\infty$ *and tends to* 0 *as* $s\uparrow\infty$ *in* Σ.

(d) *If* $s,t\in\Sigma$, *then* $s+t\in\Sigma$ *and* $\gamma(s+t)=\gamma(s)\gamma(t)$.

(e) $\gamma(t)=1-G(t)$ *for* $t\in\Sigma$.

(f) *For each* n, $\pi\in\Pi$, *random variables* τ_n^π; $\Omega\to\Sigma$ *for almost everywhere.*

Associated with F_x in Assumption 2(2), we use the following notations for a transition law p and a reward function r:

(i) $\bar r(x,a):=r(x,a)F_x(G)\in M(SA)$

(ii) $\bar p_a(x,dy):=p_a(x;dy)F_x(\gamma)\in P(S|S)$ for $a\in A$

(iii) $\bar r(f)(x_0,\cdots,x_n;dy):=r(f)(x_0,\cdots,x_n)F_{x_n}(G)\in M(S^{n+1})$

(iv) $\bar p_f(x_0,\cdots,x_n;dy):=p_f(x_0,\cdots,x_n;dy)F_{x_n}(\gamma)\in P(S|S^{n+1})$

for a mapping $f:S^{n+1}\to A$, where p_a, $r(f)$, p_f are defined in the paragraph below Definition 2.3.

3. Definition of an optimal policy

For a policy $\pi\in\Pi$ in the semi-Markov decision process (S,A,p_0,p,r,F,G), processes $X^\pi(t)$, $A^\pi(t)$, $R_\pi(t)$, $t\in R_+$ are defined in section 2, which designate the state of the system, the utilizing mapping of the policy and the total reward respectively. In this section we shall consider the expectation of the *total reward* $R_\pi(0)$ which called the expected total reward for the policy π.

Let, for $\pi\in\Pi$,

$$I(\pi):=E[R_\pi(0)],$$

$$I^*:=\sup_{\pi\in\Pi}I(\pi).$$

That is, $I(\pi)$ is the *expected total reward starting at time* 0 and I^* is the *maximal expected reward.*

DEFINITION 3. 1. An optimal policy and an ε-optimal policy are defined as follows;

(i) a policy $\pi^*\in\Pi$ is an *optimal policy* iff $I(\pi^*)=I^*$,

(ii) a policy $\pi^*\in\Pi$ is an ε-*optimal policy* iff $I(\pi^*)\geqq I^*-\varepsilon$ for $\varepsilon>0$.

This section deals with the existence of an ε-optimal and an equivalent version of the optimality. The existence of an ε-optimal Markov policy and an ε-optimal stationary policy for any $\varepsilon>0$ are argued in section 4 and that of an optimal stationary policy is in section 5. In section 6 we shall show the properties of the optimal stationary policy.

THEOREM 3. 1. *For any* $\varepsilon>0$, *there exists an* ε-*optimal policy.*

PROOF. Since r is bounded and G is a measure on $[0,\infty]$, $I(\pi)\leqq\|r\|$ for any $\pi\in\Pi$. Hence $I^*=\sup I(\pi)<\infty$. This follows immediately, for any $\varepsilon>0$, there is a policy $\pi\in\Pi$ such that $I(\pi)\geqq I^*-\varepsilon$, that is, an ε-optimal policy.

In order to state the equivalent version of the optimality, we prepare some notations and lemmas of which subsequent sections are in need.

For a policy $\pi=(f_n;n\geqq 1)$, set $R_\pi(t)$, $t\geqq 0$ be the total reward and random variables τ_n^π, σ-algebras $\mathcal{G}_n^\pi$, notations $\bar r(f_n)$, $\bar p_{f_n}$ are defined for the reward function r and the transition law p in section 2. Let define $I_n(\pi)$, $J_{n\pi}$ such that

$$I_n(\pi):=E[R_\pi(\tau_n^\pi)\,|\,\mathcal{G}_n^\pi] \quad (n\geqq 0),$$

$$J_{n\pi}(x_0,\cdots,x_{n-1}):=\bar r(f_n)(x_0,\cdots,x_{n-1})$$
$$+\sum_{k=0}^\infty\{\bar p_{f_n}\cdots\bar p_{f_{n+k}}\bar r(f_{n+k+1})\}(x_0,\cdots,x_{n-1}) \quad (n\geqq 1).$$

Clearly $I_n(\pi)$ is a random variable which is $\mathcal{G}_n^\pi$-measurable and has a finite expectation for each n. Since r is bounded, $J_{n\pi}\in M(S^n)$ for all π. Moreover let

$$J_n(x_0, \cdots, x_{n-1}) := \sup_{\pi\in\Pi} J_{n\pi}(x_0, \cdots, x_{n-1}),$$

$$T_0 := T_{0\pi} := \{x_0 \in S;\ p_0\{x_0\} = P(X^\pi(0) = x_0) > 0\},$$

$$T_{n\pi} := \{(x_0, \cdots, x_n) \in S^{n+1};\ (x_0, \cdots, x_{n-1}) \in T_{n-1,\pi} \text{ and}$$

$$p_{f_n}(x_0, \cdots, x_{n-1}; \{x_n\}) > 0\} \quad \text{for} \quad n \geqq 1.$$

If a policy π is Markov or stationary, we write J_π instead of $J_{1\pi}$.

LEMMA 3. 2. *Let v_{k+1} be bounded functions with $v_{k+1} \in M(S^{k+1})$, $k \geqq 0$. For a policy $\pi = (f_n;\ n \geqq 1)$, it holds that*

(a) $E[v_{k+1}(Z_0^\pi, \cdots, Z_k^\pi)\{G(\tau_{k+1}^\pi) - G(\tau_k^\pi)\} \mid \mathcal{G}_k^\pi] = \gamma(\tau_k^\pi)\bar{v}_{k+1}(Z_0^\pi, \cdots, Z_k^\pi)$ $(k \geqq 0)$,

(b) *for* $k \geqq n+1$, $n \geqq 0$,

$E[v_{k+1}(Z_0^\pi, \cdots, Z_k^\pi)\{G(\tau_{k+1}^\pi) - G(\tau_k^\pi)\} \mid \mathcal{G}_n^\pi] = \gamma(\tau_n^\pi)\bar{p}_{f_{n+1}} \cdots \bar{p}_{f_k}\bar{v}_{k+1}(Z_0^\pi, \cdots, Z_n^\pi)$

where $\bar{v}_{k+1}(x_0, \cdots, x_k) := v_{k+1}(x_0, \cdots, x_k)F_{x_k}(\gamma)$.

PROOF. (a) From the definition of σ-algebra $\mathcal{G}_n^\pi$, $Z_0^\pi, \cdots, Z_n^\pi$, τ_n^π are $\mathcal{G}_n^\pi$-measurable. Hence

$$E[v_{n+1}(Z_0^\pi, \cdots, Z_n^\pi)\{G(\tau_{n+1}^\pi) - G(\tau_n^\pi)\} \mid \mathcal{G}_n^\pi]$$
$$= v_{n+1}(Z_0^\pi, \cdots, Z_n^\pi)\{E[G(\tau_{n+1}^\pi) \mid \mathcal{G}_n^\pi] - G(\tau_n^\pi)\}.$$

The conditional expectation equals

$$E[G(\tau_{n+1}^\pi) \mid \mathcal{G}_n^\pi] = \int_{-\infty}^{\infty} G(t)F(Z_n^\pi, \tau_n^\pi; dt) = \int_0^\infty G(t + \tau_n^\pi)F_{Z_n^\pi}(dt)$$

by lemma 2.1 and Assumption 1. Using lemma 2. 4 (b) and (f), (a) is proved easily.

(b) At first we show that (b) holds for $k = n+1$. We have

$$E[v_{n+2}(Z_0^\pi, \cdots, Z_{n+1}^\pi)\{G(\tau_{n+2}^\pi) - G(\tau_{n+1}^\pi)\} \mid \mathcal{G}_{n+1}^\pi]$$
$$= \gamma(\tau_{n+1}^\pi)\bar{v}_{n+2}(Z_0^\pi, \cdots, Z_{n+1}^\pi)$$

replacing k with n in the equality (a). Since $\mathcal{G}_{n+1}^\pi \supset \mathcal{G}_n^\pi$, lemma 2. 1 (a), Assumption 2 (2) and lemma 2. 4 (d) follow that

$$E[v_{n+2}(Z_0^\pi, \cdots, Z_{n+1}^\pi)\{G(\tau_{n+2}^\pi) - G(\tau_{n+1}^\pi)\} \mid \mathcal{G}_n^\pi]$$
$$= E[\gamma(\tau_{n+1}^\pi)\bar{v}_{n+2}(Z_0^\pi, \cdots, Z_{n+1}^\pi) \mid \mathcal{G}_n^\pi]$$
$$= \int_{-\infty}^{\infty}\int_S \gamma(t)\bar{v}_{n+1}(Z_0^\pi, \cdots, Z_n^\pi, y)\,p_{f_{n+1}}(Z_0^\pi, \cdots, Z_n^\pi; dy)F(Z_n^\pi, \tau_n^\pi; dt)$$
$$= \gamma(\tau_n^\pi)\,p_{f_{n+1}}\bar{v}_{n+2}(Z_0^\pi, \cdots, Z_n^\pi)F_{Z_n^\pi}(\gamma)$$
$$= \gamma(\tau_n^\pi)\bar{p}_{f_{n+1}}\bar{v}_{n+2}(Z_0^\pi, \cdots, Z_n^\pi).$$

So we have proved (b) for $k = n+1$.

For a general $k \geqq n+1$, the fact $\mathcal{G}_{k-1}^\pi \supset \mathcal{G}_{k-2}^\pi \supset \cdots \supset \mathcal{G}_n^\pi$ leads to similar calculations. The details are omitted.

LEMMA 3. 3. (a) $J_{n\pi}(x_0, \cdots, x_{n-1}) = \bar{r}(f_n)(x_0, \cdots, x_{n-1})$

$$+ \int_S J_{n+1,\pi}(x_0, \cdots, x_{n-1}, y)\bar{p}_{f_n}(x_0, \cdots, x_{n-1}; dy)$$

(b) $I_0(\pi) = J_{1\pi}(X^\pi(0))$,

$I_n(\pi) = \gamma(\tau_n^\pi)J_{n+1,\pi}(X^\pi(0), \cdots, X^\pi(\tau_n^\pi))$ $(n \geqq 1)$

(c) $I(\pi) = p_0 J_{1\pi}$

PROOF. (a) It is clear because of the definition of $J_{n\pi}$.

(b) Lemma 2. 3 yields that

$$I_n(\pi) := E[R_\pi(\tau_n^\pi) \mid \mathcal{G}_n^\pi]$$
$$= \sum_{k=n}^{\infty} E[r(Z_n^\pi, f_{k+1}(Z_0^\pi, \cdots, Z_k^\pi))\{G(\tau_{k+1}^\pi) - G(\tau_k^\pi)\} \mid \mathcal{G}_n^\pi]$$

for $n \geqq 0$. If we apply Lemma 3. 2 for bounded functions;

$$v_{k+1}(x_0, \cdots, x_k) := r(f_{k+1})(x_0, \cdots, x_k),$$

(a) and (b) imply that

$$E[r(f_{n+1})(Z_0^\pi, \cdots, Z_n^\pi)\{G(\tau_{n+1}^\pi) - G(\tau_n^\pi)\} \mid \mathcal{G}_n^\pi]$$
$$= \gamma(\tau_n^\pi)\bar{r}(f_{n+1})(Z_0^\pi, \cdots, Z_n^\pi)$$

and

$$E[r(f_{k+1})(Z_0^\pi, \cdots, Z_k^\pi)\{G(\tau_{k+1}^\pi) - G(\tau_k^\pi)\} \mid \mathcal{G}_n^\pi]$$
$$= \gamma(\tau_n^\pi)\{\bar{p}_{f_{n+1}} \cdots \bar{p}_{f_k}\bar{r}(f_{k+1})\}(Z_0^\pi, \cdots, Z_n^\pi) \quad (k \geqq n+1)$$

respectively. Hence

$$I_n(\pi) = \gamma(\tau_n^\pi)\bar{r}(f_{n+1})(Z_0^\pi, \cdots, Z_n^\pi)$$
$$+ \gamma(\tau_n^\pi)\sum_{k=n+1}^{\infty}\{\bar{p}_{f_{n+1}} \cdots \bar{p}_{f_k}\bar{r}(f_{k+1})\}(Z_0^\pi, \cdots, Z_n^\pi).$$

Noting that $X^\pi(\tau_k^\pi) = Z_k^\pi$, $k \geqq 0$, we obtain

$$I_n(\pi) = \gamma(\tau_n^\pi)J_{n+1,\pi}(X^\pi(0), \cdots, X^\pi(\tau_n^\pi)) \quad (n \geqq 0).$$

Particularly, if we set $n=0$, $I_0(\pi) = J_{1\pi}(X^\pi(0))$ follows from $\tau_0^\pi = 0$ and $\gamma(0) = 1$. For (c), observe that $I(\pi) = E[R_\pi(0)] = E[E[R_\pi(0) \mid \mathcal{G}_0^\pi]] = E[I_0(\pi)] = E[J_{1\pi}(X^\pi(0))] = p_0 J_{1\pi}$ and so the proof of the lemma is complete.

Now we prove that

THEOREM 3. 4. *The following three statement are equivalent under Assumption* 1—2:

(a) π^* *is optimal,*

(b) $J_{1\pi^*}(x) = J_1(x)$ *for any* $x \in T_0$,

(c) $J_{n\pi^*}(x_0, \cdots, x_{n-1}) = J_n(x_0, \cdots, x_{n-1})$ *for any* $(x_0, \cdots, x_{n-1}) \in T_{n-1,\pi^*}$ $n \geqq 1$.

PROOF. (a) → (b): Assume $\pi^* = (f_n^*)$ to be optimal, but $J_{1\pi^*}(\tilde{x}) < J_1(\tilde{x})$ for some $\tilde{x} \in T_0$, i.e. for some $\tilde{x}$ such that $p_0(\tilde{x}) > 0$. Then there is some policy $\pi' = (f_n') \in \Pi$ for which $J_{1\pi'}(\tilde{x}) > J_{1\pi^*}(\tilde{x})$. Define a policy $\pi = (f_n)$ by

$$f_n(x,\cdots,x_{n-1}):=\begin{cases} f'_n(\tilde{x},x_1,\cdots,x_{n-1}) & \text{if } x_0=\tilde{x},\\ f^*_n(x_0,x_1,\cdots,x_{n-1}) & \text{if } x_0\neq\tilde{x}.\end{cases}$$

Obviously $\pi=(f_n)$ is a policy with

$$J_{1\pi}(y)=\begin{cases} J_{1\pi'}(\tilde{x}) & \text{if } y=\tilde{x},\\ J_{1\pi^*}(y) & \text{if } y\neq\tilde{x}.\end{cases}$$

We get

$$I^*=I(\pi^*)<\int_S J_{1\pi}(y)\,p_0(dy)=I(\pi)\leqq I^*$$

which is a contradiction.

(b)→(c): The proof goes by induction on n. Statement (c) is true for $n=1$, because this is exactly statement (b). Now we assume $\pi^*=(f^*_n)$ to be

$$J_{n\pi^*}(x_0,\cdots,x_{n-1})=J_n(x_0,\cdots,x_{n-1})$$

for any $(x_0,\cdots,x_{n-1})\in T_{n-1,\pi^*}$ but

$$J_{n+1,\pi^*}(\tilde{x}_0,\cdots,\tilde{x}_n)<J_{n+1}(\tilde{x}_0,\cdots,\tilde{x}_n)$$

for some $(\tilde{x}_0,\cdots,\tilde{x}_n)\in T_{n\pi^*}$.
It follows that

$$J_{n+1,\pi^*}(\tilde{x}_0,\cdots,\tilde{x}_n)<J_{n+1,\pi'}(\tilde{x}_0,\cdots,\tilde{x}_n)$$

for some policy $\pi'=(f'_n)$. Now we construct a policy $\pi=(f_n)$ by

$$f_k(x_0,\cdots,x_{k-1}):=f^*_k(x_0,\cdots,x_{k-1}) \quad \text{for any } (x_0,\cdots,x_{k-1}) \quad (1\leqq k\leqq n),$$

$$:=\begin{cases} f'_k(\tilde{x}_0,\cdots,\tilde{x}_{k-1}) & \text{if } (x_0,\cdots,x_{k-1})=(\tilde{x}_0,\cdots,\tilde{x}_{k-1})\\ f^*_k(x_0,\cdots,x_{k-1}) & \text{otherwise}\end{cases} \qquad (k\geqq n+1).$$

Obviously π is a policy with

$$J_{n+1,\pi}(x_0,\cdots,x_n)=\begin{cases} J_{n+1,\pi'}(\tilde{x}_0,\cdots,\tilde{x}_n) & \text{if } (x_0,\cdots,x_n)=(\tilde{x}_0,\cdots,\tilde{x}_n)\\ J_{n+1,\pi^*}(x_0,\cdots,x_n) & \text{otherwise.}\end{cases}$$

From lemma 3.3 (a) we obtain

$$\begin{aligned} J_n(\tilde{x}_0,\cdots,\tilde{x}_{n-1})&=J_{n\pi^*}(\tilde{x}_0,\cdots,\tilde{x}_{n-1})\\ &=\bar{r}(f^*_n)(\tilde{x}_0,\cdots,\tilde{x}_{n-1})+\int_S J_{n+1,\pi^*}(\tilde{x}_0,\cdots,\tilde{x}_{n-1},y)\,\bar{p}_{f^*_n}(\tilde{x}_0,\cdots,\tilde{x}_{n-1};dy)\\ &<\bar{r}(f_n)(\tilde{x}_0,\cdots,\tilde{x}_{n-1})+\int_S J_{n+1,\pi}(\tilde{x}_0,\cdots,\tilde{x}_{n-1},y)\,\bar{p}_{f_n}(\tilde{x}_0,\cdots,\tilde{x}_{n-1};dy)\\ &=J_{n\pi}(\tilde{x}_0,\cdots,\tilde{x}_{n-1})\leqq J_n(\tilde{x}_0,\cdots,\tilde{x}_{n-1})\end{aligned}$$

which contradicts our assumption that (c) is true for n.

(c)→(a): The definition of $J_1(x)$ yields $p_0J_1\geqq p_0J_{1\hat{\pi}}$ for any $\hat{\pi}$. Hence (c) implies

$$I(\pi^*)=p_0J_{1\pi^*}=p_0J_1\geqq p_0J_{1\hat{\pi}}=I(\hat{\pi})$$

for any $\hat{\pi}$. Since there is an ε-optimal policy, say, $\pi=(f_n)$ for any $\varepsilon>0$ by Theorem 3.1,

$$I(\pi^*)\geqq I(\pi)\geqq I^*-\varepsilon.$$

Therefore $I(\pi^*)\geqq I^*-\varepsilon$ holds for arbitrary $\varepsilon>0$. Thus we obtain $I(\pi^*)\geqq I^*$. The assertion (a) is now immediate because the alternative inequality holds trivially. This completes the proof of Theorem 3.2.

4. Policy reduction

In this section we are going to show that an ε-optimal Markov policy exists and moreover an ε-optimal stationary policy exists for any $\varepsilon>0$. These results are contained by the following Theorem 5.6 which states that an optimal stationary policy exists, but these are useful to prove the theorem.

ASSUMPTION 3. Assume that

$$\beta:=\sup_{x\in S}F_x(\gamma)<1.$$

This behaves as the so-called *discounted factor* in the ordinary Markov decision process.

LEMMA 4.1. *For policies* $\pi^1:=(f^1_n;n\geqq1)$ *and* $\pi^2:=(f^2_n;n\geqq1)$ *with* $f^1_k=f^2_k$ $(1\leqq k\leqq N)$, *it holds that*

$$|I(\pi^1)-I(\pi^2)|\leqq2\|r\|\frac{\beta^N}{1-\beta}.$$

PROOF. Indeed lemma 3.3 (c) follows

$$I(\pi)=p_0\bar{r}(f_1)+\sum_{n=1}^{\infty}p_0\bar{p}_{f_1}\cdots\bar{p}_{f_n}\bar{r}(f_{n+1})$$

for any policy $\pi=(f_n;n\geqq1)$ and so

$$\begin{aligned} &I(\pi^1)-I(\pi^2)\\ &=p_0\bar{p}_{f^1_1}\bar{p}_{f^1_2}\cdots\bar{p}_{f^1_N}[\bar{r}(f^1_{N+1})(.)-\bar{r}(f^2_{N+1})(.)\\ &\quad+\sum_{n=N+2}^{\infty}\{\bar{p}_{f^1_{N+1}}\cdots\bar{p}_{f^1_{n-1}}\bar{r}(f^1_n)\}(.)\\ &\quad-\sum_{n=N+2}^{\infty}\{\bar{p}_{f^2_{N+1}}\cdots\bar{p}_{f^2_{n-1}}\bar{r}(f^2_n)\}(.)].\end{aligned}$$

Observe that, for any $x, x_0, \cdots, x_n \in S$ and a policy $\pi=(f_n)$,

$$F_x(G) \leqq \int_S F_x(dy)=1,$$

$$|r(f_n)(x_0, \cdots, x_{n-1})| \leqq \|r\|,$$

$$|\bar{r}(f_n)(x_0, \cdots, x_{n-1})| \leqq \|r\| F_{x_{n-1}}(G) \leqq \|r\|,$$

$$\{\bar{p}_{f_{n-1}}\|\bar{r}\|\}(x_0, \cdots, x_{n-2}) \leqq \|r\| F_{x_{n-2}}(\gamma) \leqq \|r\| \quad (n \geqq 2),$$

$$\{\bar{p}_{f_{N+1}} \cdots \bar{p}_{f_{n-1}}\|\bar{r}\|\}(x_0, \cdots, x_N) \leqq \beta^{n-1-N}\|r\| \quad (n \geqq N+2).$$

Hence we obtain the result;

$$|I(\pi^1)-I(\pi^2)| \leqq \{p_0 p_{f_1^1} \cdots p_{f_N^1}\}[2 \sum_{n=N+1}^{\infty} \beta^{n-1-N}\|r\|]$$

$$=2\|r\|\frac{\beta^N}{1-\beta}.$$

Firstly we improve Theorem 3.1 by showing that ε-optimal Markov policy exists in Theorem 4.4. The assumptions for a transition law p and a reward function r are necessary.

ASSUMPTION 4. (1) The function

$$pu(x, a) := \int_S u(y) p(x, a; dy)$$

is upper semi-continuous in $a \in A$ for each $x \in S$ and $u \in M(S)$.

(2) The reward function $r=r(x, a)$ is also upper semi-continuous in $a \in A$ for each $x \in S$.

Let $\pi=(f_n) \in \Pi$ be an arbitrary Markov policy. For $a \in A$, let $(a, \pi) := (a, f_1, f_2, \cdots)$ and so $(a, \pi) \in \Pi$ is Markov. Using the Markov policy (a, π) let

$$K_\pi(x, a) := J_{1(a,\pi)}(x), \quad x \in S, \quad a \in A.$$

Then K_π is a function defined on SA with a Markov policy π.

Since

$$K_\pi(x, a)=\bar{r}(x, a)+\sum_{n=1}^{\infty}\{\bar{p}_a \bar{p}_{f_1} \cdots \bar{p}_{f_n} \bar{r}(f_{n+1})\}(x)$$

and it converge uniformly by Assumption 3.

LEMMA 4.2. *The function $K_\pi(x, .)$ is upper semi-continuous for each x and a Markov policy π.*

The next lemma is a policy improvement by a Markov policy.

LEMMA 4.3. *Suppose that a policy $\pi=(f_n; n \geqq 1)$ itself is not necessary Markov but ${}^N\pi := (f_n; n \geqq N+1)$ is Markov for some $N<\infty$. Then we can construct a Markov policy π^* such that $I(\pi^*) \geqq I(\pi)$.*

PROOF. For a policy $\pi=(f_n; n \geqq 1)$ which ${}^{n+1}\pi$ is Markov, we shall construct a policy $\pi^*=(f_n^*)$ with properties;

(a) ${}^n\pi^*$ is Markov,

(b) $I(\pi^*) \geqq I(\pi)$.

This shows lemma 4.3. Indeed, repeating this procedure from $n=N-1$ to 0, finally we attain the seeking Markov policy.

We now expose the above construction. Let f_{n+1}^* be a mapping S into A with

$$K_{(n+1)\pi}(x, f_{n+1}^*(x))=\max K_{(n+1)\pi}(x, a)$$

for all $y \in S$ where $(n+1)\pi := {}^{n+1}\pi := (f_{n+1}, f_{n+2}, \cdots)$. The maximum is taken all a in the action space A and it exists because of lemma 4.2. If we set $\pi^* := (f_1, \cdots, f_n, f_{n+1}^*, f_{n+2}^*, \cdots)$ for a given policy $\pi=(f_1, \cdots, f_n, f_{n+1}^*, f_{n+2}^*, \cdots)$, this policy π^* satisfies the required properties (a) and (b). Because ${}^{n+1}\pi^*={}^{n+1}\pi$ and that f_{n+1}^* is a mapping from S into A and so ${}^n\pi^*=(f_{n+1}^*, f_{n+2}^*, \cdots)$ is Markov. (b) is proved since

$$K_{(n+1)\pi^*}(f_{n+1}^*)(x_n) := K_{(n+1)\pi^*}(x_n, f_{n+1}^*(x_n))$$

$$=K_{(n+1)\pi}(x_n, f_{n+1}^*(x_n)) \geqq K_{(n+1)\pi}(x_n, f_{n+1}(x_0, \cdots, x_n))$$

$$=: K_{(n+1)\pi}(f_{n+1})(x_0, \cdots, x_n) \quad \text{for } x_0, \cdots, x_n \in S \text{ and}$$

$$I(\pi^*)-I(\pi)$$

$$=p_0 p_{f_1} \cdots p_{f_n}\{K_{(n+1)\pi^*}(f_{n+1}^*)(.)-K_{(n+1)\pi}(f_{n+1})(.)\} \geqq 0.$$

THEOREM 4.4. *For any $\varepsilon>0$, there exists an ε-optimal Markov policy under Assumption 1—4.*

PROOF. Let $\pi=(f_n)$ be an $\frac{\varepsilon}{2}$-optimal policy, that is, $I^* \leqq I(\pi)+\frac{\varepsilon}{2}$, which exists by Theorem 3.1. Let N represents an integer which satisfies $2\|r\|\frac{\beta^N}{1-\beta}<\frac{\varepsilon}{2}$. As $\beta<1$ by Assumption 3, it is sufficient to select N so that β^N is small.

Define a policy $\tilde{\pi}=(\tilde{f}_n)$, using the $\frac{\varepsilon}{2}$-optimal policy π and N, as follows:

for any $x_0, \cdots, x_{n-1} \in S$,

$$\tilde{f}_n(x_0, \cdots, x_{n-1}) := \begin{cases} f_n(x_0, \cdots, x_{n-1}) & \text{if } n \leqq N, \\ g_n(x_{n-1}) & \text{if } n \geqq N+1 \end{cases}$$

where $(g_n; n \geqq N+1)$ is an arbitrary sequence of mappings from S into A.

Since ${}^N\tilde{\pi}=(g_n; n \geqq N+1)$ is a Markov policy, there is a Markov policy π^* with $I(\pi^*) \geqq I(\pi)$ by lemma 4.3. Also it holds $I(\pi) \leqq I(\tilde{\pi})+\frac{\varepsilon}{2}$ because that lemma 4.1 follows

$$|I(\pi)-I(\tilde{\pi})| \leqq 2\|r\|\frac{\beta^N}{1-\beta}<\frac{\varepsilon}{2}.$$

Hence $I(\pi)\leqq I(\pi^*)+\frac{\varepsilon}{2}$. It now follows immediately

$$I^*\leqq I(\pi)+\frac{\varepsilon}{2}\leqq I(\pi^*)+\varepsilon.$$

The proof of Theorem 4.4 is complete.

Secondarily we improve Theorem 4.4 by showing in Theorem 4.10 that an ε-optimal stationary policy exists. Two operators L_f, U are defined for the preparation.

DEFINITION 4.1. Let $f; S\to A$ be a mapping. For $u\in M(S)$, let $L_f u$ be an element of $M(S)$ whose value at $x\in S$ is

$$L_f u(x):=\bar{r}(f)(x)+\bar{p}_f u(x).$$

If $f(x)$ equals a for all $x\in S$ then we write $L_a:=L_f$.

Let U be an operator on $M(S)$ whose value at $x\in S$ is

$$Uu(x):=\max_{a\in A} L_a u(x)$$

for $u\in M(S)$. Note that, for each $x\in S$ and $u\in M(S)$, $L_a u(x)$ is upper semi-continuous in $a\in A$ and so the operator is well defined.

Associated with each mapping $f; S\to A$ is a corresponding operator L_f, mapping $M(S)$ into $M(S)$. $L_f u$ is our expected income, as a function of the initial state, if we start using decision f but are terminated at the beginning of the second jump with a final reward $u(x)$, where x is the state at termination. $L_f^n:=L_f(L_f^{n-1})$ has a similar interpretation, replacing "second" by "$n+1^{st}$". The following interpretation of U will be justified later. $U^n u$, a function of the initial state, is our optimal expected retern over all Markov policies if we start using an optimal policy but are terminated at the beginning of the $n+1^{st}$ jump with a final reward $u(x)$ where x is the state at termination.

Here are some properties of L_f and U as the following two lemmas.

LEMMA 4.5. *Let f be a mapping from S into A.*

(a) $L_f J_{1\pi}=J_{1(f,\pi)}$ *for* $\pi\in\Pi$ *(π may not be Markov)*

(b) $L_f(u+c)(x)=L_f u(x)+cF_x(\gamma)$ *where c is a constant and* $u\in M(S)$

(c) *L_f is monotone, that is, if $u\leqq v$, then* $L_f u\leqq L_f v$

(d) *For any Markov policy $\pi=(f_n)$ and $u\in M(S)$,*

$$L_{f_1}\cdots L_{f_n}u(x):=L_{f_1}(L_{f_2}(\cdots(L_{f_n}u)))(x)$$

converge to $J_\pi(x)$ uniformly in x as $n\to\infty$.

PROOF. We shall prove only (d). If we set $u_n(x):=J_{(n)\pi}(x)$, $J_\pi=L_{f_1}\cdots L_{f_n}u_n$ and $\|u_n\|\leqq\frac{\|r\|}{1-\beta}$ from (a). Since

$$|L_f v(x)-L_f w(x)|\leqq\|v-w\|F_x(\gamma)\leqq\beta\|v-w\|$$

for a mapping $f: S\to A$ and $v, w\in M(S)$, it now follows as

$$\begin{aligned}\sup_x|L_{f_1}\cdots L_{f_n}u(x)-J_\pi(x)|\\ \leqq\sup_x L_{f_1}\cdots L_{f_n}|u-u_n|(x)\\ \leqq\beta^{n-1}\|u-u_n\|\leqq\beta^{n-1}\Big\{\|u\|+\frac{\|r\|}{1-\beta}\Big\}\end{aligned}$$

that $L_{f_1}\cdots L_{f_n}u(x)\to J_\pi(x)$ uniformly in $x\in S$.

LEMMA 4.6. (a) *U is monotone.*

(b) $U(u+c)(x)=Uu(x)+cF_x(\gamma)$ *where c is constant and* $u\in M(S)$.

(c) $L_f u(x)\leqq Uu(x)$ *for any mapping* $f: S\to A$.

(d) *U is a contraction with modulus β, that is,*

$$\|Uu-Uv\|\leqq\beta\|u-v\| \quad \textit{for} \quad u, v\in M(S).$$

(e) *U has a unique fixed point u^* in $M(S)$, that is,*

$$Uu^*=u^* \quad \textit{and} \quad \|U^n u-u^*\|\leqq\beta^n\|u-u^*\|$$

for any $u\in M(S)$, $n\geqq 1$ *where* $U^n:=U(U^{n-1})$ *iteratively.*

PROOF. (d) $\|Uu-Uv\|=\sup_x|Uu-Uv|(x)\leqq\|u-v\|\sup_x F_x(\gamma)\leqq\beta\|u-v\|$. (e) $M(S)$ is a complete metric space. Hence, from the fixed point theorem of Banach, the contraction mapping U has a unique fixed point.

A relation between the operator L_f and U is that

LEMMA 4.7. *For each $u\in M(S)$, there is a mapping $f: S\to A$ such that $L_f u=Uu$.*

PROOF. $L_a u(x)$ is upper semi-continuous in $a\in A$ for each $x\in S$. The fact that the action space A is compact yields that it attains its maximum. Let $f(x)$ be one of the point in A which attains the maximum. Then f is a mapping from S into A and satisfies the property.

DEFINITION 4.2. We say that $u^*\in M(S)$ satisfies the *optimality equation* (abbrev. *OE*) if it is a fixed point of U, that is, $u^*(x)=Uu^*(x)$ for each $x\in S$.

LEMMA 4.8. *If $u^*\in M(S)$ satisfies the OE, then there is a stationary policy f^∞ such that $J_{f^\infty}(x)=u^*(x)$ for each $x\in S$. Hence $I(f^\infty)=p_0u^*$.*

PROOF. Indeed lemma 4.7 follows that there is a mapping $f: S\to A$ which satisfies $L_f u^*(x)=Uu^*(x)$ for each $x\in S$. The fact that $L_f^n u^*=u^*$ and $L_f^n u^*\to J_{f^\infty}$ as $n\to\infty$ by lemma 4.5 (d), implies $J_{f^\infty}=u^*$. The later equality is immediate because $I(f^\infty)=p_0 J_{f^\infty}$.

LEMMA 4.9. *If $u^*\in M(S)$ satisfies the OE, then, for any Markov policy π,*

(a) $u^*(x)\geqq J_\pi(x)$, $x\in S$,

(b) $p_0u^*\geqq I(\pi)$.

PROOF. (a) Let $\pi=(f_n)$ be any Markov policy. For each element f_n of π, $L_{f_n}u^*(x)\leqq Uu^*(x)=u^*(x)$, $x\in S$ by lemma 4.6 (c) and so $L_{f_1}\cdots L_{f_n}u^*(x)\leqq u^*(x)$, $x\in S$. Letting $n\to\infty$ we obtain the assertion (a). (b) is clear if we integrate the both side of (a) by the distribution p_0.

Now we state the following theorem but will be improved in Theorem 5.6.

THEOREM 4. 10. *For any $\varepsilon>0$, there exists an ε-optimal stationary policy under Assumption* 1—4.

PROOF. Since there is u^* satisfying the OE, we have from lemma 4.8 and 4.9 that there exists an stationary policy f^∞ with $I(f^\infty)=p_0u^*\geqq I(\pi)$ for any Markov π. Let π^* be the ε-optimal Markov policy in Theorem 4.4. Then

$$I(f^\infty)\geqq I(\pi^*)\geqq I^*-\varepsilon.$$

This completes the proof.

5. Optimality equation and optimal stationary policy

We shall state the existance of an optimal stationary policy and the relation between the optimality equation and the optimal reward in Theorem 5.6 and 5.7.

LEMMA 5. 1. *Let $\varepsilon>0$ and $u\in M(S)$.*

(a) *If $L_fu(x)-\varepsilon\leqq u(x)$, $x\in S$ for some mapping f, then*

$$J_{f^\infty}(x)-\frac{\varepsilon}{1-\beta}\leqq u(x),\quad x\in S\quad \text{and so}\quad I(f^\infty)-\frac{\varepsilon}{1-\beta}\leqq p_0u$$

(b) *If $L_fu(x)+\varepsilon\geqq u(x)$, $x\in S$ for some mapping f, then*

$$J_{f^\infty}(x)+\frac{\varepsilon}{1-\beta}\geqq u(x),\quad x\in S\quad \text{and so}\quad I(f^\infty)+\frac{\varepsilon}{1-\beta}\geqq p_0u$$

(c) *If $L_au(x)-\varepsilon\leqq u(x)$, $x\in S$ for all $a\in A$, then*

$$J_\pi(x)-\frac{\varepsilon}{1-\beta}\leqq u(x),\quad x\in S\quad \text{for any Markov and so}\quad I(\pi)-\frac{\varepsilon}{1-\beta}\leqq p_0u.$$

(d) *If $L_au(x)+\varepsilon\geqq u(x)$, $x\in S$ for all $a\in A$. then*

$$J_\pi(x)+\frac{\varepsilon}{1-\beta}\geqq u(x),\quad x\in S\quad \text{for any Markov and so}\quad I(\pi)+\frac{\varepsilon}{1-\beta}\geqq p_0u.$$

PROOF. Only (c) is proved and others are omitted. Let $\pi=(f_n)$ be an arbitraly Markov policy. The condition yields that $L_{f_n}u(x)-\varepsilon\leqq u(x)$, $x\in S$ for all n. By induction on n we obtain $L_{f_1}\cdots L_{f_n}u(x)\leqq u(x)+\varepsilon(1+\beta+\cdots+\beta^{n-1})$. Letting $n\to\infty$ and the integration of both side by p_0 completes the proof.

The following lemma is useful as the policy improvement.

LEMMA 5. 2. *If $L_fJ_\pi\geqq J_\pi$ on T_0 for some policy π, then $J_{f^\infty}\geqq J_\pi$ on T_0 and so $I(f^\infty)\geqq I(\pi)$.*

LEMMA 5. 3. *If a Markov policy π^* satisfies $L_aJ_{\pi^*}\leqq J_{\pi^*}$ for all $a\in A$, that is, $UJ_{\pi^*}\leqq J_{\pi^*}$, then $J_{\pi^*}\geqq J_\pi$ for any Markov π.*

PROOF. It is clear from lemm 5. 1 (c) with letting $u(x)=J_{\pi^*}(x)$, $x\in S$ and $\varepsilon=0$.

LEMMA 5. 4. *A function $u\in M(S)$ satisfies $L_fu(x)=u(x)$, $x\in S$ for a mapping $f:S\to A$ iff $u(x)=J_{f^\infty}(x)$ with $f^\infty:=(f,f,\cdots)$.*

PROOF. If $L_fu(x)=u(x)$, $x\in S$ then $|L_fu(x)-u(x)|\leqq\varepsilon$, $x\in S$ for any $\varepsilon>0$. Lemma 5.1 (a), (b) imply that $|J_{f^\infty}(x)-u(x)|\leqq\frac{\varepsilon}{1-\beta}$, $x\in S$ for any $\varepsilon>0$. Letting $\varepsilon\to0$, it must be that $J_{f^\infty}(x)=u(x)$, $x\in S$. The converse is immediate because $u(x)=J_{f^\infty}(x)=L_fJ_{f^\infty}(x)=L_fu(x)$.

LEMMA 5. 5. *If $u^*(x)$, $x\in S$ satisfies the OE, then $I^*=p_0u^*$, that is p_0u^* equals the maximum expected reward.*

PROOF. 1) First we show $p_0u^*\geqq I^*$. Theorem 4.4 implies that for any $\varepsilon>0$, there is an Markov policy π^* with $I(\pi^*)\geqq I^*-\varepsilon$. Since u^* satisfies the OE and π^* is Markov, it follows $p_0u^*\geqq I(\pi^*)$ by lemma 4.9 (b). Therefore $p_0u^*\geqq I^*-\varepsilon$. Letting $\varepsilon\to0$, we obtain $p_0u^*\geqq I^*$.

2) By lemma 4.8, there is a stationary policy f^∞ such that $I(f^\infty)=p_0u^*$. It is immediate that $p_0u^*\leqq I^*$ because

$$p_0u^*\leqq\sup_fI(f^\infty)\leqq\sup_\pi I(\pi)=I^*,$$

where the supremum of f is taken over those mapping $f:S\to A$ and that of π is over all policies.

Combining 1) and 2) completes the proof.

Now we assert our main results. The first is the existence of an optimal stationary policy and the second is the relation between the OE and the optimal reward.

THEOREM 5. 6. *There exists an optimal stationary policy under Assumption* 1—4.

PROOF. Let $u^*(x)$, $x\in S$ be the solution of the OE. Since there is a stationary policy f such that $I(f^\infty)=p_0u^*$ by lemma 4.8 and p_0u^* is the maximum expected reward by lemma 5.5, consequently $I(f^\infty)=I^*$. This is nothing but to say that f^∞ is the optimal stationary policy. Hence the theorem is proved.

THEOREM 5. 7. (a) *If $J_{1\pi^*}(x)$, $x\in S$ with some policy π^* satisfies the OE, then the policy π^* is optimal.*

(b) *Conversely if π^* is the optimal policy and if $T_0=S$, then $J_{1\pi^*}(x)$, $x\in S$ satisfies the OE.*

PROOF. (a) is immediate consequence of lemma 5.5. (b) Let $u^*(x)$, $x\in S$ be the solution of the OE. There is a stationary policy f^∞ such that $u^*=J_{f^\infty}$ by lemma 4.8 and the proof of Theorem 5.6 implies that the stationary policy f^∞ is optimal. Hence $J_{f^\infty}(x)=J_1(x)$ and so $u^*(x)=J_{1\pi^*}(x)$ for $x\in S$. This completes the proof.

6. Properties of optimal stationary policy

Suppose the following Assumption 5 holds with Assumption 1—4. Then the optimal stationary policy in section 5 has the properties stated in Theorem 6.2. Specially the implication (a) of Theorem 6.2 represents the principle of optimality in the semi-Markov decision process.

ASSUMPTION 5. For the distribution function F_x, $x \in S$, there are a subset Σ_x of Σ and a right continuous function ϕ_x such that

(i) $F_x(s+t)=\phi_x(t)F_x(s)+F_x(t)$, $s, t \in \Sigma_x$,

(ii) for each $t \geqq 0$, $\phi_x(t)=\phi_x(t_0)$, $x \in S$ where $t_0 := \inf\{s \in \Sigma_x;\ t<s\}$.

Under Assumption 5 the next lemma holds, which is the basic recursive relation.

LEMMA 6. 1. *If a policy π is stationary, then*

$$E[R_\pi(t) \mid \mathcal{F}_t^\pi]=\gamma(t)J_\pi(X_t^\pi) \quad \textit{for} \quad t \in R_+.$$

PROOF. If $\Lambda \in \mathcal{F}_t^\pi$, then for each n there is a set $\Lambda_n \in \mathcal{G}_n^\pi$ such that

$$\Lambda \cap \{\tau_n^\pi \leqq t < \tau_{n+1}^\pi\} = \Lambda_n \cap \{t<\tau_{n+1}^\pi\}$$

Therefore it is sufficient to calculate

$$E[R_\pi(t);\ \Lambda_n \cap \{t<\tau_{n+1}^\pi\}]$$

in place of $E[R_\pi(t);\Lambda]$.

First we show three assertions;

(a) $E[1;\ \tau_{n+1}^\pi>t \mid \mathcal{G}_n^\pi]=\int_{t-\tau_n^\pi}^\infty F_{Z_n^\pi}(ds)=\phi_{Z_n^\pi}(t-\tau_n^\pi)$,

(b) $E\left[\int_t^{\tau_{n+1}^\pi} r(X^\pi(s), A^\pi(s))G(dx);\ \Lambda_n \cap \{t<\tau_{n+1}^\pi\}\right]$

$$=E\left[r(Z_n^\pi, f_n(Z_n^\pi))\gamma(t)\int_{t-\tau_n^\pi}^\infty G(s+\tau_n^\pi-t)F_{Z_n^\pi}(ds);\ \Lambda_n\right],\ \Lambda_n \in \mathcal{G}_n^\pi,$$

(c) $E\left[\int_{\tau_k^\pi}^{\tau_{k+1}^\pi} r(X^\pi(s), A^\pi(s))G(ds);\ \Lambda_n \cap \{t<\tau_{n+1}^\pi\}\right]$

$$=E\left[\gamma(\tau_n^\pi)\int_{t-\tau_n^\pi}^\infty \gamma(s)F_{Z_n^\pi}(ds)\{p_{f_n}\bar{p}_{f_{n+1}}\cdots \bar{p}_{f_k}\bar{r}(f_k)\}(Z_n^\pi);\ \Lambda_n\right]$$

for $k \geqq n+1$, where a policy $\pi=(f_n)$ is stationary with $f_n=f$ for all n.

Indeed (a) is from Assumption 5 (ii). (b) follows according to the calculation;

$$E\left[\int_t^{\tau_{n+1}^\pi} r(X^\pi(s), A^\pi(s))G(ds);\ \Lambda_n \cap \{t<\tau_{n+1}^\pi\}\right]$$

$$=E[r(Z_n^\pi, f_n(Z_n^\pi))\{G(\tau_{n+1}^\pi)-G(t)\};\ \Lambda_n \cap \{t<\tau_{n+1}^\pi\}]$$

$$=E\left[r(Z_n^\pi, f_n(Z_n^\pi))\int_{t-\tau_n^\pi}^\infty F_{Z_n^\pi}(ds);\ \Lambda_n\right]$$

$$-G(t)E\left[r(Z_n^\pi, f_n(Z_n^\pi))\int_{t-\tau_n^\pi}^\infty G(s+\tau_n^\pi-t)F_{Z_n^\pi}(ds);\ \Lambda_n\right].$$

For (c), noting

$$\Lambda_n \cap \{t<\tau_{n+1}^\pi\} \in \mathcal{G}_{n+1}^\pi,$$

lemma 3.2 (b) and lemma 2.1 (a) imply that

$$E[\gamma(\tau_{n+1}^\pi)\{\bar{p}_{f_{n+1}}\cdots \bar{p}_{f_{k-1}}\bar{r}(f_k)\}(Z_{n+1}^\pi);\ \Lambda_n \cap \{t<\tau_{n+1}^\pi\}]$$

$$=E\left[\gamma(\tau_{n+1}^\pi)\int_{t-\tau_n^\pi}^\infty \gamma(s)F_{Z_n^\pi}(ds)\{p_{f_n}\bar{p}_{f_{n+1}}\cdots \bar{p}_{f_{k-1}}\bar{r}(f_k)\}(Z_n^\pi);\ \Lambda_n\right].$$

It is now immediate because of the definition J_π and above (a), (b) and (c) that

$$E[R_\pi(t);\ \Lambda_n \cap \{t<\tau_{n+1}^\pi\}]$$

$$=E[\gamma(t)\phi_{Z_n^\pi}(t-\tau_n^\pi)\bar{r}(Z_n^\pi, f_n(Z_n^\pi))$$

$$+\gamma(t)\phi_{Z_n^\pi}(t-\tau_n^\pi)\sum_{k=n+1}^\infty \{\bar{p}_{f_n}\cdots \bar{p}_{f_{k-1}}\bar{r}(f_k)\}(Z_n^\pi);\ \Lambda_n]$$

$$=\gamma(t)E[\phi_{Z_n^\pi}(t-\tau_n^\pi)J_\pi(Z_n^\pi);\ \Lambda_n]$$

$$=\gamma(t)E[J_\pi(X_t^\pi);\ \Lambda_n \cap \{t<\tau_{n+1}^\pi\}].$$

This proves the lemma.

THEOREM 6. 2. *Let π^* be the optimal stationary policy obtained in section 5. Then we have*

(a) *for any stationary $\pi \in \Pi$ and $t \geqq 0$,*

$$P(E[R_{\pi^*}(t) \mid \mathcal{F}_t^{\pi^*}] \geqq E[R_\pi(t) \mid \mathcal{F}_t^\pi] \mid X_t^{\pi^*}=X_t^\pi)=1.$$

(b) *If $X_t^{\pi^*}$, X_t^π have the same distribution and $T_0=S$, then for any stationary policy π and $t \geqq 0$,*

$$E[R_{\pi^*}(t)] \geqq E[R_\pi(t)].$$

PROOF. Both of (a) and (b) are proved by applying lemma 6.1 and the results of Thorem 3. 4 (a), (b).

References

[1] R. BELLMAN, Dynamic Programming (Princeton Univ. Press, Princeton, 1957).

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

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

[4] D. BLACKWELL, *Positive dynamic programming*, Proc. of the fifth Berkeley Symp. 1965 Vol. **I** (1967), 415–418.

[5] R.M. BLUMENTHAL and R.K. GETOOR, Markov Processes and Potential Theory (Academic Press, New York, 1968).

[6] G. DE LEVE, Generalized Markovian Decision Processes I Model and Method; II Probabilistic Background (Mathematical Centre Tracts 3 and 4, Amsterdam, 1964).

[7] G. DE LEVE, H.C. TIJMS and P.J. WEEDA, Generalized Markovian Decision Processes, Applications (Mathematical Centre Tracts 5, Amsterdam, 1970).

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

[9] K. HINDERER, Foundation of Non-stationary Dynamic Programming with Discrete Time Parameter (Springer, Berlin, 1970).

[10] A. HORDIJK, Dynamic Programming and Markov Potential Theory (Mathematisch Centrum, Amsterdam, 1974).

[11] R.A. HOWARD, Dynamic Programming and Markov Processes (Technology Press and Wiley, Massachusetts, 1960).

[12] R.A. HOWARD, *Research in semi-Markovian decision structure*, J. Operat. Res. Soc. Japan **6** (1964), 163-199.

[13] W.S. JEWELL, *Markov renewal programming* I *and* II, Operat. Res. **2** (1963), 938-971.

[14] P. KAKUMANU, *Continuously discounted Markov decision model with countable state space and action space*, Ann. Math. Statist. **42** (1971), 919-926.

[15] S.A. LIPPMAN, *Maximal average-reward policies for semi-Markov decision processes with arbitrary state and action space*, Ann. Math. Statist. **42** (1971), 1717-1726.

[16] S.A. LIPPMAN, *Semi-Markov decision processes with unbounded rewards*, Management Science **19** (1973), 717-731.

[17] A. MAITRA, *Dynamic programming for countable state systems*, Sankhya **27A** (1965), 241-248.

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

[19] A. MARTINLÖF, *Optimal control of a continuous time Markov chain with periodic transition probabilities*, Operation Research **15** (1967), 872-881.

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

[21] S.M. ROSS, *Average cost semi-Markov decision processes*, J. Appl. Probability **7** (1970), 649-656.

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

[23] A.F. VEINOTT, *On finding optimal policies in discrete dynamic programming with no discounting*, Ann. Math. Statist. **37** (1966), 1284-1294.

[24] A.F. VEINOTT, *Discrete dynamic programming with sensitive optimality criteria*, Ann. Math. Statist. **40** (1969), 1635-1660.