DISCOUNTED MARKOV DECISION PROCESSES WITH GENERAL UTILITY FUNCTIONS

Yoshinobu KADOTA
Faculty of Education, Wakayama University, Wakayama 640, Japan
yoshi-k@math.edu.wakayama-u.ac.jp

Masami KURANO
Faculty of Education, Chiba University, Chiba 263, Japan
kurano@cue.e.chiba-u.ac.jp

and

Masami YASUDA
Faculty of Science, Chiba University, Chiba 263, Japan
yasuda@math.s.chiba-u.ac.jp

ABSTRACT

We consider a maximization of the expected utility of the total discounted rewards in countable state Markov decision processes. Specifying the class of distribution functions for the present value and using its weak compactness, we established the optimality equation under a general utility. Also a g-optimal policy is constructed. As an application of g-optimality, we discuss the moment optimality introduced by Jaquette[7].

1. Introduction

A utility optimization of Markov decision processes(MDP's) with countable state and compact action spaces is considered. As for utility functions, an exponential one has many attractive properties. For example, it has a constant local risk aversion and an invariant risk premium with respect to the wealth(Fishburn[5], Pratt[11]). Several authors analyzed MDP's with exponential utility functions. Howard and Matheson[6] studied the case of finite states and actions in N horizon times. Letting N tend to infinity, they gave the policy improvement to find the policy that maximizes the time-average equivalent returns of MDP's. Chung and Sobel[2] considered the maximization of the expected utility of the total discount return random variable (called the present value) for finite MDP's and derived the optimality equations, by which an optimal policy was constructed. Porteus[10] and Denardo and Rothblum[3] dealt with the problem from the other points of view.

In this paper, we consider the same problems as those treated in Chung and Sobel[2] when the utility function is a general, in particular, continuous function, by which the practical applications will be enlarged. The method of analysis employed here is closely related to the one in Sobel[13], White[14], where the distribution function of

the present value is characterized by iterative formula and the fixed point theory. Specifying the class of distribution functions of the present value as a weak-compact space, we derive the optimality equation under a continuous utility function g, from which a g-optimal policy is constructed. In the case of the exponential utility, the optimal equation derived here is the same as that in Chung and Sobel[2]. As an application of our results, we treat the moment optimality introduced by Jaquette[7] and show that there exists a stationary policy which is moment optimal for countable state MDP's.

In Section 2, we shall prepare the several notations and define the problem to be examined. Also, the weak-compactness of distribution functions of the present value is described referring to Borkar's excellent book[1].

2. Preliminaries

We consider standard Markov decision processes specified by $(S, \{A(i)\}_{i\in S}, q, r)$, where $S = \{1, 2, \cdots\}$ denotes the set of the states of the processes, $A(i)$ is the set of actions available at each state $i \in S$, $q = (q_{ij}(a))$ is the matrix of transition probabilities satisfying that $\sum_{j\in S} q_{ij}(a) = 1$ for all $i \in S$ and $a \in A(i)$, and $r(i, a, j)$ is an immediate reward function defined on $\{(i, a, j)|i \in S, a \in A(i), j \in S\}$.

Throughout this paper, the following assumptions will be remained operative:

(i) For each $i \in S$, $A(i)$ is a closed set of a compact metric space.

(ii) For each $i, j \in S$, both $q_{ij}(\cdot)$ and $r(i, \cdot, j)$ is continuous on $A(i)$.

(iii) The function r is uniformly bounded, i.e., $0 \leq r(i, a, j) \leq M$ for all $i, j \in S$ and $a \in A(i)$.

The sample space is the product space $\Omega = (S \times A)^\infty$ such that the projection X_t, Δ_t on the t-th factors S, A describe the state and the action of t-time of the process($t \cdot \geq 0$). A policy $\pi = (\pi_0, \pi_1, \cdots)$ is a sequence of conditional probabilities π_t such that $\pi_t(A(i_t)|i_0, a_0, \cdots, i_t) = 1$ for all histories $(i_0, a_0, \cdots, i_t) \in (S \times A)^t \times S$. The set of all policies is denoted by Π. A policy $\pi = (\pi_0, \pi_1, \cdots)$ is called stationary if there exists a function f with $f(i) \in A(i)$ for all $i \in S$ such that $\pi_t(\{f(i)\}|i_0, a_0, \cdots, i_t = i) = 1$ for all $t \geq 0$ and $(i_0, a_0, \cdots, i_t) \in (S \times A)^t \times S$. Such a policy is denoted by f^∞. Let $H_t = (X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t)$ for $t \geq 0$. We assume that for each $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$,

$$P_\pi(X_{t+1} = j|H_{t-1}, \Delta_{t-1}, X_t = i, \Delta_t = a) = q_{ij}(a)$$

for all $t \geq 0, i, j \in S, a \in A(i)$. For any Borel measurable set X, $\mathcal{P}(X)$ denotes the set of all probability measures on X. Then, any initial probability measure $\nu \in \mathcal{P}(S)$ and policy $\pi \in \Pi$ determine the probability measure $P_\pi^\nu \in \mathcal{P}(\Omega)$ by a usual way.

Lemma 2.1.(e.g. see Borkar[1]) *For each $\nu \in \mathcal{P}$, $\overline{\varphi}(\nu) := \{P_\pi^\nu \in \mathcal{P}(\Omega)|\pi \in \Pi\}$ is compact in the weak topology.*

The discounted present value of the state-action process $\{X_t, \Delta_t; t = 0, 1, 2, \cdots\}$ is defined by

$$\mathcal{B} := \sum_{t=0}^{\infty} \beta^t r(X_t, \Delta_t, X_{t+1}), \tag{2.1}$$

where $\beta (0 < \beta < 1)$ is a discount factor. Let $u := M/(1-\beta)$. Then, for each $\nu \in P(S)$ and $\pi \in \Pi$, $\mathcal{B}$ is a random variable from the probability space (Ω, P_π^ν) into the interval $[0, u]$. We denote by $C[0, u]$ the set of all bounded continuous functions on $[0, u]$. Let $g \in C[0, u]$ be arbitrary. Then, interpreting this g as a utility function, our problem is to maximize the expected utility $E_\pi^\nu(g(\mathcal{B}))$ over all policies $\pi \in \Pi$, where E_π^ν is the expectation with respect to P_π^ν.

In order to analyze the above problem, it is convenient to rewrite $E_\pi^\nu(g(\mathcal{B}))$ by using the distribution function of $\mathcal{B}$ corresponding to P_π^ν. Let, for each $\nu \in P(S)$ and $\pi \in \Pi$,

$$F_\pi^\nu(z) := P_\pi^\nu(\mathcal{B} \leq z), \tag{2.2}$$

$$\Phi(\nu) := \{F_\pi^\nu(\cdot) | \pi \in \Pi\}. \tag{2.3}$$

Noting that we can identify $\Phi(\nu)$ with $\overline{\varphi}(\nu)$, the next results follows from Lemma 2.1.

Corollary 2.1. *For any $\nu \in P(S)$, $\Phi(\nu)$ is weak-compact.*

For any $g \in C[0, u]$ and $\nu \in P(S)$, we say that $\pi^* \in \Pi$ is (ν, g)-optimal if $E_{\pi^*}^\nu(g(\mathcal{B})) \geq E_\pi^\nu(g(\mathcal{B}))$ for all $\pi \in \Pi$. When π^* is (ν, g)-optimal for all $\nu \in P(S)$, π^* is simply called g-optimal.

3. g-optimality

In this section we derive the optimality equation under arbitrary continuous function g, which construct a g-optimal policy. By weak-compactness of $\Phi(\nu)$ given in Corollary 2.1. the following existence theorem holds.

Theorem 3.1. *For any $\nu \in P(S)$ and $g \in C[0, u]$, there exists a (ν, g)-optimal policy.*

Proof. By Corollary 2.1 it follows that

$$\sup_{\pi \in \Pi} E_\pi^\nu(g(\mathcal{B})) = \sup_{F \in \Phi(\nu)} \int g(z) F(dz) = \int g(z) F^*(dz)$$

for some $F^* \in \Phi(\nu)$. Corresponding to F^*, let π^* be its associated policy. Then, clearly π^* is (ν, g)-optimal. □

In order to describe the optimal equation in Theorem 3.2 below, the following lemma is useful. The proof of it is easily done from the uniform continuity of g on $[0, u]$ and Corollary 2.1.

Lemma 3.1. *For any $g \in C[0, u]$, $\int g(s + \beta z) F(dz)$ is continuous as a function of (s, F) on $[0, M] \times \Phi(\nu)$.*

For simplicity of the notation, let

$$U_t\{g\}(s,i,a,j) := \max_{F\in\Phi(j)} \int g(s+\beta^t r(i,a,j)+\beta^{t+1}z)F(dz) \tag{3.1}$$

for $t \geq 0, g \in C[0,u], s \in [0,M], i,j \in S$ and $a \in A(i)$, where if $\nu \in P(S)$ is degenerate at $\{j\}$, ν is simply denoted by j and $\Phi(\nu)$ by $\Phi(j)$. Note that by Lemma 3.1 the maximum in Eq.(3.1) is attained. Now, we can state one of our main results, which gives a necessary condition for (ν,g)-optimality.

Theorem 3.2. *For any $\nu \in P(S)$ and $g \in C[0,u]$, let $\pi^* \in \Pi$ be (ν,g)-optimal. Then for each $t \geq 0$, the following optimal equation holds.*

$$E^\nu_{\pi^*}(g(\mathcal{B})) = E^\nu_{\pi^*}\{\max_{a\in A(X_t)} \sum_{j\in S} q_{X_t j}(a) U_t\{g\}(\mathcal{B}_{t-1}, X_t, a, j)\}, \tag{3.2}$$

where $\mathcal{B}_{-1} := 0$ and $\mathcal{B}_t := \sum_{k=0}^t \beta^k r(X_k, \Delta_k, X_{k+1})$ for $t \geq 0$.

Proof. For simplicity, we denote $E^\nu_{\pi^*}$ by E. For any $\omega := (i_0, a_0, i_1, a_1, \cdots) \in \Omega$, let $\theta_t(\omega) := (i_t, a_t, i_{t+1}, \cdots)$ be a shift operator for $t \geq 1$. From the Markov property of the transition probabilities,

$$\begin{aligned} E(g(\mathcal{B})) &= E\{E\{g(\mathcal{B}_{t-1} + \beta^t r(X_t, \Delta_t, X_{t+1}) + \beta^{t+1}\mathcal{B}(\theta_{t+1}(\omega))|H_{t+1}\}\} \\ &\leq E\{E\{U_t\{g\}(\mathcal{B}_{t-1}, X_t, \Delta_t, X_{t+1})|H_t\}\} \\ &\leq E\left\{\max_{a\in A(X_t)} \sum_{j\in S} q_{X_t j}(a)\{U_t\{g\}(\mathcal{B}_{t-1}, X_t, a, j)\}\right\}. \end{aligned}$$

Since π^* is (ν,g)-optimal, the above inequalities can be all replaced by equalities. □

In order to give a sufficient condition for g-optimality, we define the sequence $\{A^*_t\}_{t=0}^\infty$ by

$$A^*_t(s,i) := \arg\max_{a\in A(i)} \sum_{j\in S} q_{ij}(a)\{U_t\{g\}(s,i,a,j)\}, \tag{3.3}$$

where for any function $h(x)$ on X,

$$\begin{aligned} &\arg\max_{x\in X}(\arg\min_{x\in X})\, h(x) \\ &\quad := \{\, x' \in X \mid x' \text{ maximizes (minimizes) } h(x) \text{ in all } x \in X\}. \end{aligned}$$

Theorem 3.3. *For any $\nu \in P(S)$ and $g \in C[0,u]$, the following (i) and (ii) hold.*
(i) *Let $\pi^* = (\pi^*_0, \pi^*_1, \cdots)$ be any (ν,g)-optimal, then*

$$P^\nu_{\pi^*}(\Delta_t \in A^*_t(\mathcal{B}_{t-1}, X_t)) = 1 \quad \text{for all} \quad t \geq 0.$$

(ii) *Let $\pi^* = (\pi^*_0, \pi^*_1, \cdots)$ be any policy satisfying*

$$\pi^*_t(A^*_t(\mathcal{B}_{t-1}, X_t)|H_t) = 1 \text{ for all } H_t \text{ and } t \geq 0.$$

Then, π^* is g-optimal.

Proof. By observing the proof of Theorem 3.2, we see that (i) holds. To prove (ii), let $\pi^t := (\pi_0^*, \pi_1^*, \cdots, \pi_t^*, \pi'_{t+1}, \pi'_{t+2}, \cdots)$, for π^* given in the statement of Theorem 3.3(ii), where $(\pi'_{t+1}, \pi'_{t+2}, \cdots)$ is a policy corresponding to $F' \in \Phi(X_{t+1})$ which satisfies the relation such that

$$\int g(\mathcal{B}_{t-1} + \beta^t r(X_t, \Delta_t, X_{t+1}) + \beta^{t+1} z) F'(dz) = U_t\{g\}(\mathcal{B}_{t-1}, X_t, \Delta_t, X_{t+1})$$

for $t \geq 0$. The policy $(\pi'_{t+1}, \pi'_{t+2}, \cdots)$ only depends on H_t. Now we shall show inductively that π^t is g-optimal for all $t \geq 0$. Let $\pi = (\pi_0, \pi_1, \cdots)$ be any policy. Then, we have

$$\begin{aligned} E_\pi^i(g(\mathcal{B})) &= E_\pi^i[E_\pi^i\{g(r(X_0, \Delta_0, X_1) + \beta\mathcal{B}(\theta_1(\omega))|H_1\}] \\ &\leq E_\pi^i\left[\max_{F\in\Phi(X_1)} \int g(r(X_0, \Delta_0, X_1) + \beta z) F(dz)\right] \\ &\leq E_{\pi^0}^i(g(\mathcal{B})) \quad \text{for all} \quad i \in S. \end{aligned}$$

Therefore, π^0 is (i, g)-optimal for all $i \in S$, that is, g-optimal. Moreover, $E_{\pi^0}^i[g(\mathcal{B})] \leq E_{\pi^0}^i[E_{\pi'}^{X_1}(g(\mathcal{B}))]$, by applying the case of $t = 0$ to $g(r(X_0, \Delta_0, X_1) + \beta x)$, where $\pi' = (\pi_1^*(H_1), \pi'_2, \pi'_3, \cdots)$. Since π^0 is g-optimal, π^1 is also so. Repeating the above argument, we can prove that π^t is g-optimal for all $t \geq 1$. Since g is uniformly continuous in $[0, u]$ for any $\epsilon > 0$, there exists $T \geq 1$ satisfying

$$|g(x + \beta^T z_1) - g(x + \beta^T z_2)| \leq \epsilon$$

for any x, z_1, z_2 such that $x + \beta^T z_j \in [0, u]$ for $j = 1, 2$. For this T, clearly it holds that

$$|E_{\pi^T}^i(g(\mathcal{B})) - E_{\pi^*}^i(g(\mathcal{B}))| \leq E_{\pi^*}^i[\sup_{z_1, z_2} |g(\mathcal{B}_T + \beta^T z_1) - g(\mathcal{B}_T + \beta^T z_2)|] \leq \epsilon$$

where the sup is taken over the range : $\mathcal{B}_T + \beta^T z_j \in [0, u]$ for $j = 1, 2$. For any $T \geq 1$, π^T is optimal, so that by $T \to \infty$ and $\epsilon \to 0$ in the above, we observe π^* is g-optimal. □

Remark 1. Consider the case when a decision maker has a linear utility function, i.e., $g(x) = x$. Then, Eq.(3.1) becomes

$$U_t\{x\}(s, i, a, j) = s + \beta^t\{r(i, a, j) + \beta \max_{F\in\Phi(j)} \int z F(dz)\},$$

and so $A_t^*(s, i)$ in Eq.(3.3) reduces

$$A_t^*(s, i) = \arg\max_{a\in A(i)} \sum_{j\in S} q_{ij}(a)\{r(i, a, j) + \beta \max_{F\in\Phi(j)} \int z F(dz)\}, \tag{3.4}$$

which is independent of s. This gives the set of optimal actions for usual MDP's under the expected total discounted reward criterion(e.g. see Ross[12]).

Remark 2. Consider the exponential utility case, i.e., $g(x) = -\exp(-\lambda x)$. Then, Eq.(3.1) becomes

$$U_t\{-e^{-\lambda x}\}(s,i,a,j) = -e^{-\lambda s}e^{-\lambda\beta^t r(i,a,j)} \min_{F\in\Phi(j)} \int \exp\{-\lambda\beta^{t+1}z\}F(dz).$$

Thus, by Eq.(3.3), we have

$$\begin{aligned} A_t^*(s,i) = \; & \arg\min_{a\in A(i)} \textstyle\sum_{j\in S} q_{ij}(a)\exp\{-\lambda\beta^t r(i,a,j)\} \\ & \times \min_{F\in\Phi(j)} \int \exp\{-\lambda\beta^{t+1}z\}F(dz). \end{aligned} \tag{3.5}$$

Observing Eq.(3.5), we note that the policy π^* constructed by Theorem 3.3 is the same as that obtained in Theorem 4 of Chung and Sobel[2], which is called λ-optimal.

4. Moment optimality

Jaquette[7] introduced the moment optimality and proved the existence of moment optimal stationary policy for finite MDP's by analyzing the negative of the Laplace transformation of $\mathcal{B}$, which is corresponding to the case of the exponential utility $g(x) = -\exp(-\lambda x)$. Here, we shall prove the existence theorem by applying the results in the proceeding section to the restricted MDP's iteratively, whose ideas were appearing in Kurano[8], Mandel[9].

First let us give several notations necessary in our discussion. For any $i \in S$ and $\pi \in \Pi$, let

$$N_n(i,\pi) := (-1)^{n+1}E_\pi^i(\mathcal{B}^n) \quad \text{for} \quad n \geq 1.$$

Let $\boldsymbol{u} = (u_1, u_2, \cdots)^t$ and $\boldsymbol{v} = (v_1, v_2, \cdots)^t$ be two vectors, where $\boldsymbol{u}^t$ denotes the transpose of $\boldsymbol{u}$. Then we write $\boldsymbol{u} \geq \boldsymbol{v}$ if $u_i \geq v_i$ for all $i = 1, 2, \cdots$. Let $N(i,\pi) := (N_n(i,\pi); n = 1, 2, \cdots)$ be a row vector and $N(\pi) := (N(i,\pi); i = 1, 2, \cdots)^t$ be an infinite matrix. For any integer $l \geq 1$ and $\pi, \pi' \in \Pi$, we write $N(i,\pi) \succeq_l N(i,\pi')$, if there is an integer $k(1 \leq k \leq l)$ such that $N_n(i,\pi) = N_n(i,\pi')$ for $1 \leq n < k$ and $N_k(i,\pi) \geq N_k(i,\pi')$. we also write $N(\pi) \succeq_l N(\pi')$ if $N(i,\pi) \succeq_l N(i,\pi')$ for any $i = 1, 2, \cdots, l$. We say that π^* is l-moment optimal if $N(\pi^*) \succeq_l N(\pi')$ for all $\pi' \in \Pi$ and that π^* is moment optimal if it is l-moment optimal for all $l = 1, 2, \cdots$ (see Jaquette[7] for details).

Let $N_1^*(i) := \max_{\pi\in\Pi} N_1(i,\pi)$. Then, applying Eq.(3.2) for $t = 0$, we obtain

$$N_1^*(i) = \max_{a\in A(i)} \sum_{j\in S} q_{ij}(a)\left(r(i,a,j) + \beta N_1^*(j)\right) \tag{4.1}$$

for all $i \in S$. Also, noting Remark 1 in section 3, let

$$A_1(i) := \arg\max_{a\in A(i)} \sum_{j\in S} q_{ij}(a)\left(r(i,a,j) + \beta N_1^*(j)\right). \tag{4.2}$$

It is easily verified that $A_1(i)$ is compact for each $i \in S$. So we denote by MDP $(S, A_1(i), q, r)$ the MDP's specified by $S, \{A_1(i); i \in S\}, q$ and r. And define $\Phi_1(i)$ for MDP $(S, A_1(i), q, r)$ similar as $\Phi(i)$ for MDP $(S, A(i), q, r)$.

By Theorem 3.3(ii), it holds that

$$N_1^*(i) = \int zF(dz) \quad \text{for all } F \in \Phi_1(i). \tag{4.3}$$

Next we define $N_2^*(i) := \min_{F\in\Phi_1(i)} \int z^2 F(dz)$. The following theorem concerning with the second moment will be established.

Theorem 4.1. (i) $N_2^*(i)$ *satisfies*

$$N_2^*(i) = \min_{a\in A_1(i)} \sum_{j\in S} q_{ij}(a) \left(\theta_2(i,a,j) + \beta^2 N_2^*(j)\right), \tag{4.4}$$

where $\theta_2(i,a,j) := r(i,a,j)^2 + 2\beta r(i,a,j) N_1^*(j)$.

(ii) *Let*

$$A_2(i) := \arg\min_{a\in A_1(i)} \sum_{j\in S} q_{ij}(a) \left(\theta_2(i,a,j) + \beta^2 N_2^*(j)\right) \tag{4.5}$$

and f^∞ *be any stationary policy with* $f(i) \in A_2(i)$ *for all* $i \in S$. *Then,* f^∞ *is 2-moment optimal.*

Proof. Letting $g(x) = -x^2$, we apply Eq.(3.2) for $t = 0$ to MDP $(S, A_1(i), q, r)$. The assertion (i) could be proved immediately. To prove (ii), we apply the results of Theorem 3.3, in which $A_t^*(s,i)$ in Eq.(3.3) is given as follows: Let $F \in \Phi_1(j)$, $a \in A_1(i)$ and $h(x) = x^2$. We have

$$\begin{aligned}
&\int h(s + \beta^t r(i,a,j) + \beta^{t+1} z) F(dz) \\
&= s^2 + 2s\beta^t \left(r(i,a,j) + \beta \int zF(dz)\right) + \beta^{2t} \int \left(r(i,a,j) + \beta z\right)^2 F(dz) \\
&= s^2 + 2s\beta^t \left(r(i,a,j) + \beta N_1^*(j)\right) + \beta^{2t} \left(\theta_2(i,a,j) + \beta^2 \int z^2 F(dz)\right).
\end{aligned}$$

So, by Eq.(4.1) and Eq.(4.2) we get

$$\begin{aligned}
& \min_{a\in A_1(i)} \textstyle\sum_j q_{ij}(a) U_t\{x^2\}(s,i,a,j) \\
= \quad & s^2 + 2s\beta^t N_1^*(i) + \beta^{2t} \min_{a\in A_1(i)} \textstyle\sum_j q_{ij}(a) \left(\theta_2(i,a,j) + \beta^2 N_2^*(j)\right).
\end{aligned}$$

Thus, we see by Eq.(3.3) that $A_t^*(s,i) = A_2(i)$ for all $t \geq 0$. From Theorem 3.3, the stationary policy f^∞ given in (ii) is shown to be 2-moment optimal. □

Applying the idea of Theorem 4.1, we can get the further results.

Theorem 4.2. *There exists a moment optimal stationary policy.*

Proof. Using $\Phi_1, N_1^*, N_2^*, A_2$ given in Eq.(4.1) – Eq.(4.6), define θ_m, Φ_{m-1}, N_m^* and A_m inductively for $m \geq 3$ as follows:

(i) Define $\Phi_{m-1}(i)$ for MDP $(S, A_{m-1}(i), q, r)$ as similar as $\Phi(i)$ for MDP $(S, A(i), q, r)$.

(ii) $N_m^*(i) := (-1)^{m+1} \max_{F\in\Phi_{m-1}(i)} \int (-1)^{m+1} x^m F(dx)$.

(iii) $\theta_m(i,a,j) := \sum_{k=0}^{m-1} \binom{m}{k} \beta^k r(i,a,j)^{m-k} N_k^*(j)$, where $N_0^*(j) = 1$ for all $j \in S$.

(iv) $A_m(i) := \arg\max_{a \in A_{m-1}(i)} (-1)^{m+1} \left[\sum_j q_{ij}(a) \left(\theta_m(i,a,j) + \beta^m N_m^*(j)\right)\right]$

Let f^∞ be any stationary policy such that $f(i) \in \bigcap_{m=0}^{\infty} A_m(i)$ for all $i \in S$. Then, it is shown analogous to the proof of Theorem 4.1 that f^∞ is l-moment optimal for all $l \geq 3$. □

References

1. V.S. Borkar, Topics in Controlled Markov Chains, Longman Scientific Technical, 1991.
2. K.J. Chung and M.J. Sobel, Discounted MDP's: Distribution functions and exponential utility maximization, *SIAM J. Control and Optimization*, **25**(1987), 49-62.
3. E.V. Denardo and U.G. Rothblum, Optimal stopping, exponential utility and linear programming, *Math. Prog.*, **16** (1979), 228–244.
4. S.N. Ethier and T.G. Kurtz, Markov Processes, Characterization and Convergence, John Wiley & Sons, New York, 1986.
5. P.C. Fishburn, Utility Theory for Decision Making, John Wiley & Sons, New York, 1970.
6. R.S. Howard and J.E. Matheson, Risk-sensitive Markov decision processes, *Manag. Sci.*, **8** (1972), 356–369.
7. S.C. Jaquette, Markov decision processes with a new optimality criterion: Discrete time, *Ann. Stat.*, **1**(1973), 496–505.
8. M. Kurano, Markov decision processes with a minimum-variance criterion, *J. Math. Anal. Appl.*, **123** (1987), 572–583.
9. P. Mandl, On the variance of controlled Markov chains, *Kybernetika*, **7** (1971), 1–12.
10. E.L. Porteus, On the optimality of structured policies in countable stage decision processes, *Manag. Sci.*, **22** (1975), 148–157.
11. J.W. Pratt, Risk aversion in the small and in the large, *Econometrica*, **32** (1964) 122–136.
12. S.M. Ross, Applied Probability Models with Optimization Applications, Holden-Day, 1970.
13. M.J. Sobel, The variance of discounted Markov decision processes, *J. Appl. Prob.*, **19** (1982) 794–802.
14. D.J. White, Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.*, **173** (1993) 634–646.

APORS'94 (July 26—29)
Fukuoka Japan

Discounted Markov decision processes with general utility functions

Yoshinobu KADOTA, Masami KURANO and Masami YASUDA
(門田 良信) (蔵野 正美) (安田 正実)
*Wakayama University** *Chiba University* *Chiba University*
*Address: 930, Sakaedani, Wakayama 640, Japan
Phone: +81-734-54-0361
E-mail: yoshi-k@edu.wakayama-u.ac.jp

Abstract

We consider the maximization of the expected utility of the present value which is a random variable of the total discounted rewards. The utility g is a general, in particular, arbitrary continuous function. The rewards are generated by Markov decision processes(MDP's) with a countable state space and with bounded reward functions.

Specifying the class of distribution functions of the present value as a weak-compact space, we show that there exists a g-optimal policy for any initial distribution. We also derive the optimality equation which constructs the g-optimal policy. If g is reduced to an exponential function, the similar results are found by Chung and Sobel[SIAM J. Ctrl. and Opt. 25(1987), p49-62] for finite MDP's. As an application of the equation, we treat the moment optimality introduced by Jaquette[Ann. Statist. 1(1973), p572-583], showing that there exists a stationary policy which is moment optimal for our case.

Key words: discounted Markov decision process, expected utility, optimality equation, g-optimality, moment optimal policy.

PROCEEDINGS OF APORS' 94

Proceedings of the
Third Conference of the Association of
Asian-Pacific Operational Research Societies (APORS)
within IFORS

Fukuoka, Japan, July 26–29 1994

Theme of the Conference

Development in Diversity and Harmony

Editors

Masanori Fushimi

Univ. Tokyo, Japan

Kaoru Tone

Saitama Univ., Japan

World Scientific

Singapore • New Jersey • London • Hong Kong