Regret-Optimality Equation in Semi-MDP's with an Absorbing Set

和歌山大·教育　門田良信
(Yoshinobu KADOTA, Faculty of Education, Wakayama University)
千葉大·教育　蔵野正美
(Masami KURANO, Faculty of Education, Chiba University))
千葉大·理学　安田正實
(Masami YASUDA, Faculty of Science, Chiba University)

Abstract. In terms of countable state Semi-Markov decision processes, the expected regret-utility incurred until the first-passage time into the absorbing set is considered. The utility of regret is represented using two variables, one is the target value and the other is the present value. We call it the regret-utility function. In order to characterize the regret-optimal policy, we derive the optimality equation, for which the uniqueness of solution is proved. As application, a few examples of regret-utility functions are given, under which some analyses are developed.

Keywords: Regret-optimal policy, Semi-Markov decision processes, General regret-utility, Optimality equation.

1. Introduction and notation

In social life or in business, the decision making is commonly executed by making the regret incurred from the decision as small as possible, where the regret means the difference between the target value and the real payoff. In this paper, the above problem will be considered in terms of countable state Semi-Markov decision processes (SMDP's) with an absorbing set.

The utility of regret is represented by a function using two variables, one is the target value and the other is the real payoff, called regret-utility function, and the problem to be solved is to minimize the expected regret-utility incurred until the first-passage time into the absorbing set. As for utility functions, refer to [3, 9] and for general utility treatment of Markov decision processes refer to [1, 2, 5, 6, 7].

In the remainder of this section, we define the regret-utility optimization problem for SMDP's to be examined in the sequel.

SMDP's are specified by (i) a countable state space $S = \{0, 1, 2, \cdots\}$, (ii) a finite action space A, (iii) transition probability matrices $p = \{p_{ij}(a) | i, j \in S, a \in A\}$, (iv) distribution functions $\{F_{ij}(\cdot|a) | i, j \in S, a \in A\}$ of the time between transitions, (v) an immediate reward r and reward rate d which are functions from $S \times A$ to $\boldsymbol{R}_+$, where $\boldsymbol{R}_+ = [0, \infty)$. When the system is in state $i \in S$ and action $a \in A$ is taken, then it moves to a new state $j \in S$ with the sojourn time τ, and the reward $r(i, a) + d(i, a)\tau$ is obtained, where the new state j and the sojourn time τ are distributed with $p_{i\cdot}(a)$ and $F_{ij}(\cdot|a)$ respectively. This process is repeated from the new state $j \in S$.

The sample space is the product space $\Omega = (S \times A \times \boldsymbol{R}_+)^\infty$. Let X_n, Δ_n and τ_{n+1} be random quantities such that $X_n(\omega) = x_n$, $\Delta_n(\omega) = a_n$ and $\tau_{n+1}(\omega) = t_{n+1}$ for all $\omega = (x_0, a_0, t_1, x_1, a_1, t_2, \cdots) \in \Omega$ and $n = 0, 1, 2, \cdots$. A policy $\pi = (\pi_0, \pi_1, \cdots)$ is a sequence of conditional probabilities π_n such that $\pi_n(A \mid x_0, a_0, t_1, \cdots, x_n) = 1$ for all histories $(x_0, a_0, t_1, \cdots, x_n) \in (S \times A \times \boldsymbol{R}_+)^n \times S$. 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 $f : S \to A$ such that $\pi_n(\{f(x_n)\} \mid x_0, a_0, t_1, \cdots, x_n) = 1$ for all $n \geqq 0$ and $(x_0, a_0, t_1, \cdots, x_n) \in (S \times A \times \boldsymbol{R}_+)^n \times S$. Such a policy is denoted by f^∞.

For any $\pi \in \Pi$, we assume that

(i) $Prob(X_{n+1} = j |\, X_0, \Delta_0, \tau_1, \cdots, X_n = i, \Delta_n = a) = p_{ij}(a)$

and

(ii) $Prob(\tau_{n+1} \leq t |\, X_0, \Delta_0, \tau_1, \cdots, X_n = i, \Delta_n = a, X_{n+1} = j) = F_{ij}(t|a)$

for all $n \geqq 0$, $i, j \in S$ and $a \in A$. Then, any initial state $i \in S$ and policy $\pi \in \Pi$ determine the probability measure $P_\pi(\cdot \mid X_0 = i)$ on Ω by a usual way.

We make the general assumption: There exists an absorbing set $J_0 \subset S$ and $J_0 \neq S$, for which that $\sum_{j \in J_0} p_{ij}(a) = 1$ and $r(i, a) = d(i, a) = 0$ for all $i \in J_0$ and $a \in A$. Let $J = S - J_0$ and N be the first-passage epoch into J_0 i.e.,

$$N = \min\{n |\, X_n \in J_0,\ n \geqq 0\}, \quad \text{where} \quad \min \emptyset = \infty.$$

The present value and the total lapsed time of the process $\{X_n, \Delta_n, \tau_{n+1} : n = 0, 1, 2, \cdots\}$ until the ℓ-th epoch are defined respectively by

$$\begin{aligned} \widetilde{D}_\ell &= \sum_{n=0}^{\ell-1} (r(X_n, \Delta_n) + \tau_{n+1} d(X_n, \Delta_n)) \quad \text{and} \\ \widetilde{\tau}_\ell &= \sum_{n=1}^{\ell} \tau_n, \quad (\, \ell \geqq 1\,). \end{aligned}$$

Let $G : \boldsymbol{R}_+ \times \boldsymbol{R}_+ \to \boldsymbol{R}$ be a Borel-measurable function, which will be called a regret-utility function. Then, for a constant g^*, called a target value, our problem is to minimize the expected regret-utility with a target g^*

$$E_\pi \left(G(g^* \widetilde{\tau}_N, \widetilde{D}_N) \mid X_0 = i \right) \quad \text{over all } \pi \in \Pi,$$

where $E_\pi(\cdot \mid X_0 = i\,)$ is the expectation with respect to $P_\pi(\cdot \mid X_0 = i\,)$. We say that $\pi^* \in \Pi$ is regret-optimal with a target g^* if

$$E_{\pi^*} \left(G(g^* \widetilde{\tau}_N, \widetilde{D}_N) \mid X_0 = i \right) \leq E_\pi \left(G(g^* \widetilde{\tau}_N, \widetilde{D}_N) \mid X_0 = i \right)$$

for all $\pi \in \Pi$ and $i \in S$.

In Section 2, under some reasonable assumptions concerning the speed with which the decision process is driven into J_0, we give the optimality equation in order to characterize the regret-optimal policy. Also, uniqueness of solution to the optimality equation is proved.

In Section 3, as applications of our results, some examples of regret-utility functions are given, under which some analyses are developed.

2. Regret-optimality and related optimality equations

To develop our discussion, the following assumption is needed.

Assumption 1. The following (i)–(ii) holds:

(i) $0 \leqq r(i, a) \leqq M_1 < \infty$, $0 \leqq d(i, a) \leqq M_2 < \infty$ for all $i \in S$, $a \in A$ and some M_1 and M_2.

(ii) There exist $L > 0$, $B > 0$ with

$$L \leqq \int_0^\infty tF_{ij}(dt|\, a) \leqq B \quad \text{for all } i,\ j \in S \ \text{ and } \ a \in A.$$

For each $i \in J$ and $n \geqq 0$, we define $e_i(n)$ by

$$e_i(n) = \sup_{\pi \in \Pi} P_\pi(X_n \in J \,|\, X_0 = i\,),$$

which means the maximal probability of being not yet absorbed in J_0 at the n-th epoch. Putting $e(n) = \sup_{i \in J} e_i(n)$, it clearly holds (cf. [4]) that $e(n+1) \leqq e(n)$ and $e(m+n) \leqq e(m)e(n)$ for all m, $n \geqq 0$. The following assumption is needed.

Assumption 2. It holds that $\delta_0 := \sum_{n=0}^{\infty} e(n) < \infty$.

If the following Assumption 2' holds, Assumption 2 follows.

Assumption 2'. There exist $0 < \eta_0 < 1$ and $n_0 \geqq 1$ such that $e(n_0) < 1 - \eta_0$.

In fact, if Assumption 2' holds, we have that

$$\begin{aligned} \delta_0 &= \sum_{n=0}^{\infty} e(n) = \sum_{k=0}^{\infty} \sum_{n=0}^{n_0 - 1} e(kn_0 + n) \leqq \sum_{k=0}^{\infty} n_0 e(kn_0) \\ &\leqq n_0 \sum_{k=0}^{\infty} e(n_0)^k \leqq n_0 \eta_0^{-1} < \infty, \end{aligned}$$

which shows that Assumption 2 holds.

Since $P_\pi(N > n|\, X_0 = i\,) \leqq e(n)$ for $n \geqq 0$, it holds that $E_\pi(N|\, X_0 = i\,) \leqq \delta_0$, which implies $\lim_{n\to\infty} nP_\pi(N > n|\, X_0 = i\,) = 0$ for any $\pi \in \Pi$.

Here, we define an optimal value function when starting from the initial state $i \in S$ by

$$g_i(c_1,\, c_2) = \inf_{\pi \in \Pi} E_\pi \left(G(c_1 + g^* \widetilde{\tau}_N,\ c_2 + \widetilde{D}_N) \,|\, X_0 = i \right). \tag{2.1}$$

By the above definition, we observe that $g_i(c_1, c_2) = G(c_1, c_2)$ for $i \in J_0$ and $g_i(0, 0)$ is the optimal expected regret-utility in our optimization problem.

The following assumption is needed to characterize the optimal value function.

Assumption 3. There exists a $K > 0$ such that

$$\int_0^\infty \left| G(c_1 + g^* t,\ c_2 + r(i,\, a) + d(i,\, a)t) - G(c_1,\, c_2) \right| F_{ij}(dt|\, a) \leqq K \tag{2.2}$$

for all c_1, $c_2 \in \boldsymbol{R}_+$, i, $j \in S$ and $a \in A$.

Remark. If $G(c_1, c_2)$ is differentiable and $\left|\dfrac{\partial G(c_1, c_2)}{\partial c_1}\right|$ and $\left|\dfrac{\partial G(c_1, c_2)}{\partial c_2}\right|$ are uniformly bounded in $(c_1, c_2) \in \boldsymbol{R}_+ \times \boldsymbol{R}_+$, Assumption 3 holds from applying the mean value theorem and Assumption 1. Hereafter, Assumption 1, 2 and 3 will be remained operative.

Lemma 2.1. *There exists a* $K^* > 0$ *such that*

$$\int_0^\infty \left|g_i(c_1 + g^*t,\ c_2 + r(i, a) + d(i, a)t) - g_i(c_1, c_2)\right| F_{ij}(dt|\, a) \leqq K^* \tag{2.3}$$

for all $c_1,\ c_2 \in \boldsymbol{R}_+,\ \ i,\ j \in S$ *and* $a \in A$.

Proof. Let t be such that $g_i(c_1 + g^*t,\ c_2 + r(i, a) + d(i, a)t) - g_i(c_1, c_2) \geqq 0$ and $\varepsilon > 0$. Then, by (2.1) there exists a policy $\pi = \pi\{i, a, t\}$ (depending on i, a, t) satisfying that

$$\begin{aligned}
&|g_i(c_1 + g^*t,\ c_2 + r(i, a) + d(i, a)t) - g_i(c_1, c_2)\,| \\
&\leqq E_\pi\Big(G(c_1 + g^*t + g^*\widetilde{\tau}_N,\ c_2 + r(i, a) + d(i, a)t + \widetilde{D}_N) - G(c_1 + g^*\widetilde{\tau}_N,\ c_2 + \widetilde{D}_N)\Big) + \varepsilon \\
&= \sum_{n=0}^{\infty} P(N = n)E\Big[E\Big[G\big(c_1' + g^*\tau_n,\ c_2' + r(X_{n-1}, \Delta_{n-1}) + d(X_{n-1}, \Delta_{n-1})\tau_n\big) \\
&-G\big(c_1'' + g^*\tau_n,\ c_2'' + r(X_{n-1}, \Delta_{n-1}) + d(X_{n-1}, \Delta_{n-1})\tau_n\big)\,\Big|\, N = n,\ H_n\Big]\Big] + \varepsilon,
\end{aligned}$$

where $c_1' = c_1 + g^*t + g^*\widetilde{\tau}_{n-1},\quad c_2' = c_2 + r(i, a) + d(i, a)t + \widetilde{D}_{n-1},$
$c_1'' = c_1 + g^*\widetilde{\tau}_{n-1},\quad c_2'' = c_2 + \widetilde{D}_{n-1},$
$H_n = (X_0,\ \Delta_0,\ \tau_1, \cdots, X_n),\quad P := P_\pi(\cdot\,|\,X_0 = i)$ and $E := E_\pi(\cdot\,|\,X_0 = i)$.
Applying Assumption 3, we have that

$$\begin{aligned}
&\left|g_i(c_1 + g^*t,\ c_2 + r(i, a) + d(i, a)t) - g_i(c_1, c_2)\right| \\
&\leqq \sum_{n=0}^{\infty} P(N = n)E\Big[E\Big[(G(c_1'\ c_2') - G(c_1'', c_2''))\ |\, N = n,\ H_{n-1}\Big] + 2K\Big] + \varepsilon \\
&\ \vdots \\
&\ \vdots\ \ \text{(repeating the same discussion)} \\
&\ \vdots \\
&\leqq \sum_{n=0}^{\infty} P(N = n)\Big(G(c_1 + g^*t,\ c_2 + r(i, a) + d(i, a)t) - G(c_1, c_2) + 2nK\Big) + \varepsilon \\
&= G(c_1 + g^*t,\ c_2 + r(i, a) + d(i, a)t) - G(c_1, c_2) + 2K\sum_{n=0}^{\infty} nP(N = n) + \varepsilon
\end{aligned}$$

Since $\sum_{n=0}^{\infty} nP(N = n) = \sum_{n=0}^{\infty} P(N \geqq n) \leqq \sum_{n=0}^{\infty} e(n) = \delta_0$, by letting $\varepsilon \to 0$ in the above, we get that

$$\begin{aligned}
&\left|g_i(c_1 + g^*t,\ c_2 + r(i, a) + d(i, a)t) - g_i(c_1, c_2)\right| \\
&\qquad \leqq \left|G(c_1 + g^*t,\ c_2 + r(i, a) + d(i, a)t) - G(c_1, c_2)\right| + 2K\delta_0.
\end{aligned} \tag{2.4}$$

Similarly, for any t with $g_i(c_1,\ c_2) - g_i(c_1 + g^*t,\ c_2 + r(i,\ a) + d(i,\ a)t) > 0$, we get (2.4). Therefore, applying Assumption 3 again, we obtain (2.3) with $K^* = K(2\delta_0 + 1)$. □

In the same way as the proof of Lemma 2.1, we can prove the following.

Lemma 2.2. *For any $i \in J$, and $c_1,\ c_2 \in \boldsymbol{R}_+$, it holds that*

$$g_i(c_1,\ c_2) \geqq G(c_1,\ c_2) - K\delta_0. \tag{2.5}$$

We denote by $\mathcal{B}(\boldsymbol{R}_+ \times \boldsymbol{R}_+)$ the set of all bounded Borel measurable functions on $\boldsymbol{R}_+ \times \boldsymbol{R}_+$. For any set $h = (h_i :\ i \in J)$ with $h_i \in \mathcal{B}(\boldsymbol{R}_+ \times \boldsymbol{R}_+)$, we define $U\{h\}(c_1,\ c_2 |\ i,\ a)$ by

$$\begin{aligned} U\{h\}(c_1,\ c_2|\ i,\ a) &= \sum_{j\in J} p_{ij}(a) \int_0^\infty h_j(c_1 + g^*t,\ c_2 + r(i,\ a) + d(i,\ a)t) F_{ij}(dt|a) \\ &+ \sum_{j\in J_0} p_{ij}(a) \int_0^\infty G(c_1 + g^*t,\ c_2 + r(i,\ a) + d(i,\ a)t) F_{ij}(dt|a) \end{aligned} \tag{2.6}$$

for $c_1,\ c_2 \in \boldsymbol{R}_+$, $i \in J$ and $a \in A$.

Obviously, for each $i \in J$ and $a \in A$, $U\{h\}(\cdot,\ \cdot\,|\ i,\ a) \in \mathcal{B}(\boldsymbol{R}_+ \times \boldsymbol{R}_+)$.

Here, we can state one of main results, whose proof is done by a slight modification of that of Theorem 3.2 and 3.3 in our preceding paper [6], so that the proof is omitted.

Theorem 2.1. *(i) The optimal value functions g_i, $i \in J$ satisfy the following optimality equation:*

$$g_i(c_1,\ c_2) = \min_{a\in A} U\{g\}(c_1,\ c_2|\ i,\ a) \tag{2.7}$$

for all $i \in J$, and $c_1,\ c_2 \in \boldsymbol{R}_+$, where $g = (g_i :\ i \in J)$.

(ii) Let $\pi^* = (\pi_0^*,\ \pi_1^*, \cdots) \in \Pi$ be any policy satisfying

$$\pi_n^* \left(A^*(g^*\widetilde{\tau}_n,\ \widetilde{D}_n :\ X_n) \,|\, H_n\right) = 1 \quad \text{on} \quad \{X_n \in J\} \tag{2.8}$$

for all $n \geqq 1$ and H_n, where

$$A^*(c_1, c_2 : i) = \text{argmin}_{a\in A} U\{g\}(c_1, c_2|i, a)$$

for $c_1, c_2 \in \boldsymbol{R}_+$ and $i \in J$. Then, π^* is regret-optimal with a target g^*.

The following theorem asserts the uniqueness of solution to the optimality equation (2.7). **Theorem 2.2.** *There exists a unique solution to the optimality equation (2.7) in $\boldsymbol{C}$,*

where

$$\boldsymbol{C} = \{h = (h_i :\ i \in J) |\ h_i \in \mathcal{B}(\boldsymbol{R}_+ \times \boldsymbol{R}_+) \ \ \textit{for all} \ \ i \in J \\ \textit{and } h \textit{ satisfies the statement of Lemma 2.1.}\}$$

Proof. Let $h = (h_i : i \in J)$, $h' = (h'_i : i \in J)$ be solutions to (2.7) and $h, h' \in \boldsymbol{C}$. Then, from (2.6) and (2.7), there is an $\bar{a} \in A$ such that

$$\begin{aligned} &|h_i(c_1, c_2) - h'_i(c_1, c_2)| \\ &\leqq \sum_{j \in J} p_{ij}(\bar{a}) \left| \int_0^\infty h_j(c_1 + g^* t,\ c_2 + r(i, a) + d(i, a)t) F_{ij}(dt | \bar{a}) \right. \\ (2.9) \qquad &\left. - \int_0^\infty h'_j(c_1 + g^* t,\ c_2 + r(i, a) + d(i, a)t) F_{ij}(dt | \bar{a}) \right| \\ &\leqq \sum_{j \in J} p_{ij}(\bar{a}) \left(\left| h_j(c_1, c_2) - h'_j(c_1, c_2) \right| + 2\overline{K} \right) \end{aligned}$$

for some $\overline{K} > 0$.

Repeating the relation (2.9), we get that

$$|h_i(c_1, c_2) - h'_i(c_1, c_2)| \leqq 2\overline{K} \sum_{n=0}^{\infty} e(n) = 2\overline{K}\delta_0 < \infty.$$

So, if we put $\|h_i - h'_i\| = \sup_{c_1, c_2 \in R_+} |h_i(c_1, c_2) - h'_i(c_1, c_2)|$, then $\|h_i - h'_i\| \leqq 2\overline{K}\delta_0$, and from the first inequality in (2.9), we get

$$(2.10) \qquad \|h_i - h'_i\| \leqq \sum_{j \in J} p_{ij}(\bar{a}) \|h_j - h'_j\| \quad \text{for} \quad i \in J.$$

Repeating (2.10), we obtain

$$(2.11) \qquad \|h - h'\| \leqq e(n)\, \|h - h'\| \quad \text{for all} \quad n \geqq 1,$$

where $\|h - h'\| = \sup_{i \in J} \|h_i - h'_i\|$. Letting $n \to \infty$ and noting that $e(n) \to 0$ from Assumption 2, (2.11) means $\|h - h'\| = 0$. Thus, $h = h'$, so that uniqueness of solutions follows. □

3. Examples

In the following examples, the results in the preceding section are applied to the cases of some types of regret-utility functions.

Example 1. Consider the case that $G(x, y) = x - y$. From Remark in Section 2, we observe that Assumption 3 holds. Putting

$$g_i = \inf_{\pi \in \Pi} E_\pi(g^* \widetilde{\tau}_N - \widetilde{D}_N \,|\, X_0 = i),$$

We get from (2.1) that

$$\begin{aligned} (3.1) \qquad g_i(c_1, c_2) &= \inf_{\pi \in \Pi} E_\pi(c_1 + g^* \widetilde{\tau}_N - c_2 - \widetilde{D}_N \,|\, X_0 = i) \\ &= c_1 - c_2 + g_i \qquad (\ i \in J,\ c_1, c_2 \in \boldsymbol{R}_+\ ). \end{aligned}$$

Thus, the optimality equation (2.7) becomes:

$$g_i = \min_{a\in A}\Big\{ - R(i,\, a) + \sum_{j\in J} p_{ij}(a) g_j + g^* \overline{\tau}(i,\, a)\Big\} \tag{3.2}$$

for $i \in J$, where $R(i,\, a) = r(i,\, a) + d(i,\, a)\overline{\tau}(i,\, a)$ and

$$\overline{\tau}(i,\, a) = \sum_{j\in S} p_{ij}(a) \int_0^\infty tF_{ij}(dt|a) \quad \text{for } i \in J \text{ and } a \in A.$$

Applying Theorem 2.1, we can obtain a regret-optimal policy using the unique solution of (3.2).

Remark. We consider recurrent Semi-MDP's and put:

$$J_0 = \{0\}, \quad N = \min\{n|\, X_n = 0,\, n \geqq 1\} \quad \text{and}$$

$$g^* = \sup_{\pi\in\Pi} \frac{E_\pi(\widetilde{D}_N|\, X_0 = 0)}{E_\pi(N|\, X_0 = 0)}.$$

Then, (3.2) with $g_0 = 0$ is corresponding to the optimality equation for the average case. In fact, it holds (cf. [8, 10]) that

$$\min_{a\in A}\Big\{ - R(0,\, a) + \sum_{j\neq 0} p_{0j}(a) g_j + g^* \overline{\tau}(0,\, a)\Big\} = 0,$$

so that putting $g_0 = 0$, (3.2) holds for all $i \in S$.

Example 2. Consider the case of the exponential type: $G(x,\, y) = -e^{-\lambda(x-y)}$, ($\lambda > 0$). When the target value g^* is sufficiently large such that $g^*t - r(i,\, a) - d(i,\, a)t \geqq 0$ for all $t \geqq 0$, $i \in S$ and $a \in A$, Assumption 3 in Section 2 holds obviously.

Let

$$g_i = \inf_{\pi\in\Pi} E_\pi\Big(- e^{-\lambda(g^*\widetilde{\tau}_N - \widetilde{D}_N)}\,|\, X_0 = i\Big),$$

for $i \in J$. Then, $g_i(c_1,\, c_2) = e^{-\lambda(c_1 - c_2)} g_i$, so that the optimality equation (2.7) comes to

$$g_i = \min_{a\in A}\Big[\sum_{j\in J} p_{ij}(a) R(i,\, a,\, j) g_j - \sum_{j\in J_0} p_{ij}(a) R(i,\, a,\, j)\Big]$$

where $R(i,\, a,\, j) = \int_0^\infty e^{-\lambda(g^*t - r(i,\, a) - d(i,\, a)t)} F_{ij}(dt|a)$ for $i \in J$, $j \in S$ and $a \in A$

Applying Theorem 2.1, we get a regret-optimal policy for the exponential regret-utility case.

References

[1] Chung, K. J. and Sobel, M. J. (1987), Discounted MDP's: Distribution functions and exponential utility maximization. *SIAM J. Control Optimization* **25**, 49-62.

[2] Denardo, E.V. and Rothblum, U. G. (1979), Optimal stopping, exponential utility and linear programming. *Math. Prog.* **16**, 228–244.

[3] Fishburn, P. C. (1970), Utility Theory for Decision Making. John Wiley & Sons, New York.

[4] Hinderer, K. and Waldmann, K. H. (2003), The critical discount factor for finite Markovian decision processes with an absorbing set. *Math. Mech. Oper. Res.* **57**, 1–19.

[5] Howard, R. S. and Matheson, J. E. (1972), Risk-sensitive Markov decision processes. *Management Science* **8**, 356–369.

[6] Kadota, Y., Kurano, M. and Yasuda, M. (1995), Discounted Markov decision processes with general utility functions. In *Proceeding of APORS' 94*, 330–337, World Scientific.

[7] Kadota, Y., Kurano, M. and Yasuda, M. (1998), On the general utility of discounted Markov decision processes. *Int. Trans. Oper. Res.* **5**, 27–34.

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

[9] Pratt, J. W. (1964), Risk aversion in the small and in the large. *Econometrica* **32**, 122–136.

[10] Ross, S. M. (1970), Applied Probability Models with Optimization Applications. Holden-Day.