Regret optimality in semi-Markov decision processes with an absorbing set

Yoshinobu Kadota
Faculty of Education, Wakayama University, Wakayama 640-8510, Japan
yoshi-k@math.edu.wakayama-u.ac.jp

Masami Kurano
Faculty of Education, Chiba University, Chiba 263-8522, Japan
kurano@faculty.chiba-u.jp

Masami Yasuda
Faculty of Science, Chiba University, Chiba 263-8522, Japan
yasuda@math.s.chiba-u.ac.jp

Abstract

The optimization problem of general utility case is considered for countable state semi-Markov decision processes. The regret-utility function is introduced as a function of two variables, one is a target value and the other is a present value. We consider the expectation of the regret-utility function incured until the reaching time to a given absorbing set. In order to characterize the regret optimal policy, we derive the optimality equation and then prove the uniqueness of solution. As application, two examples of regret-utility functions are used to illustrate the analysis for these models.

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

1 Introduction and notation

The optimization problem of general utility case is considered for countable state semi-Markov decision processes. If a decision maker assesses a random variable reward Y by use of a general utility function U, the following U-certainty equivalent may be considered. U-certainty equivalent means a real-valued quantity $E(U, Y)$, whose utility $U(E(U, Y))$ equals to the expectation of utility $U(Y)$, that is, it is defined by

$$U(E(U,Y)) = E[U(Y)] \ \text{ or } \ E(U,Y) = U^{-1}(E[U(Y)]).$$

The above equation means that decision maker would be indifferent about receiving between the random rewards Y and the non-random amount $E(U,Y)$. See Fishburn[9] and Pratt[16] in detail.

In our model, a performance criterion is the U-certainty equivalent of the total reward until the reaching time to the absorbing set. The study of Markov decision processes endowed with the risk-sensitive average criterion and their related works have been developing by many authors [1, 2, 3, 4, 5, 11, 20], in which the utility is exponential function with constant risk sensitivity λ, i.e., $U_\lambda(y) = sign(\lambda)e^{\lambda y}$ if $\lambda \neq 0$, $= y$ if $\lambda = 0$. In this case, the U-certainty equivalent $E(U_\lambda, Y)$ is expressed by an explicit formula:

$$E(U_\lambda, Y) = \begin{cases} \dfrac{1}{\lambda}\ln(E[e^{\lambda Y}]), & \lambda \neq 0 \\ E[Y] & \lambda = 0 \end{cases}$$

Our paper does not specify only this kind of utility U_λ but we consider the case of the U-certainty equivalent $E(U,Y)$ of a general utility function U for a random variable Y. However it is too much difficult tp express the U-certainty equivalent $E(U,Y)$ explicitly. So we will introduce a regret which evaluates the difference between the target value and the real payoff.

We assume that the utility of regret is represented by a function of 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 incured until the reaching time to the absorbing set.

In order to characterize the regret optimal policy, we derive the regret optimality equation. Then the uniqueness of solution will be proved. As application, two examples of regret-utility function are illustrated and some analysis are developed. There are many kind of variability-risk analysis for Markov Decision Processes; variance, percentile, etc, which appeared in [7, 8, 14, 18, 19, 21, 22]. Also for a general utility of Markov decision processes, refer to [5, 6, 11, 12, 13].

In the remainder of this section, we define the regret-utility optimization problem for semi-MDP's to be examined in the sequel. semi-MDP's are specified by

(i) a countable state space: $S = \{0, 1, 2, \cdots\}$,

(ii) a finite action space: $A = \{1, 2, \cdots, m\}, m < \infty$,

(iii) transition probability distributions: $\{(p_{ij}(a); j \in S) | i \in S, a \in A\}$,

(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 a 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 $n = 0,\ 1,\ 2, \cdots$. Let $H_n = (X_0, \Delta_0, \tau_1, \cdots, X_n)$ be a history until time n. A policy $\pi = (\pi_0,\ \pi_1, \cdots)$ is a sequence of conditional probabilities $\pi_n = \pi_n(\cdot \mid H_n)$ such that $\pi_n(A \mid H_n) = 1$ for all histories $H_n \in (S \times A \times \boldsymbol{R}_+)^n \times S$. 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 $\pi_n(\{f(X_n)\} \mid H_n) = 1$ for all $n \geq 0$ and $H_n \in (S \times A \times \boldsymbol{R}_+)^n \times S$. Such a policy is denoted by f^∞.

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

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

(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 $n \geq 0,\ i,\ j \in S$ 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 $J_0 \neq S$ such that $\sum_{j \in J_0} p_{ij}(a) = 1$ and $r(i, a) = d(i, a) = 0$ hold for all $i \in J_0, a \in A$. Let $J = S \setminus J_0$ and N be the reaching time to J_0, i.e., $N = \min\{n |\ X_n \in J_0,\ n \geq 0\}$, provided that $\min \emptyset = \infty$. 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 time are defined respectively by

$$\widetilde{D}_\ell = \sum_{n=0}^{\ell-1} (r(X_n,\ \Delta_n) + \tau_{n+1} d(X_n,\ \Delta_n))$$

and $\widetilde{\tau}_\ell = \sum_{n=1}^{\ell} \tau_n, \quad (\ell \geq 1)$.

Motivated from the previous discussion, we introduce the following function G which is used in the evaluation between a target value and a present value. Let $G : \boldsymbol{R}_+ \times \boldsymbol{R}_+ \to \boldsymbol{R}$ be a Borel-measurable function and call it as a regret-utility function. For a constant g^*, called as a target value, our problem is to minimize the expected regret-utility with a target g^*

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

where $E_\pi(\cdot \mid X_0 = i)$ is the expectation with respect to $P_\pi(\cdot \mid X_0 = i)$. For example, the difference between a target value g^* and an average of present value $\widetilde{D}_N / \widetilde{\tau}_N$ is evaluated by

$$G(g^* \widetilde{\tau}_N, \widetilde{D}_N) = -\exp\{-\lambda(g^* \widetilde{\tau}_N - \widetilde{D}_N)\}$$

which is analyzed in Example 2 in Section 3. This situation have related to our previous model on the general utility of Markov decision processes [12, 13]. We say that $\pi^* \in \Pi$ is regret optimal with a target g^* if

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

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, a few examples of regret-utility functions are given, under which some analysis are developed.

2 Regret optimality and related optimality equations

To develop our discussion, the following assumption is needed. These require that the process should be natural not pathological and also that reward r and its rate d are bounded.

Assumption 1. For all $i, j \in S$, $a \in A$,

(i) there exists M_1 and M_2 such that

$$0 \le r(i,a) \le M_1 < \infty, \quad 0 \le d(i,a) \le M_2 < \infty,$$

(ii) there exist $L > 0$, $B > 0$ such that $L \le \int_0^\infty t\, F_{ij}(dt|\, a) \le B$.

For each $i \in J$ and $n \geq 0$, 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 time. Putting $e(n) = \sup_{i \in J} e_i(n)$, it clearly holds(cf. [10]) that $e(n+1) \leq e(n)$ and $e(m+n) \leq e(m)e(n)$ for all $m,\, n \geq 0$.

The following assumption is needed.

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

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

In stead of Assumption 2, we could assume Assumption 2'. In fact, if Assumption 2' holds, we have that

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

which shows that Assumption 2 holds. In addition, since $P_\pi(N > n|\, X_0 = i\,) \leq e(n)$ for $n \geq 0$, it holds that $E_\pi(N|\, X_0 = i\,) \leq \delta_0$ and then it implies $\lim_{n \to \infty} nP_\pi(N > n|\, X_0 = i\,) = 0$ for any $\pi \in \Pi$. Because $kP(N > k|X_0 = i) \leq k \sum_{n>k} P(N = n|X_0 = i) \leq \sum_{n>k} nP(N = n|X_0 = i) \to 0 (k \to \infty)$.

Now we define an optimal value function starting from the initial state i and for $c_1,\, c_2 \in \boldsymbol{R}_+$ by

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

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 utilized to characterize the optimal value function.

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

$$\int_0^\infty \Big| G(\overline{c_1}, \overline{c_2}) - G(c_1,\, c_2) \Big| F_{ij}(dt|\, a) \leq K \tag{2.2}$$

where $\overline{c_1} = c_1 + g^* t, \overline{c_2} = c_2 + r(i,\, a) + d(i,\, a)t$ 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. *For any $i \in J$ and $c_1, c_2 \in \boldsymbol{R}_+$, it holds that*

$$|g_i(c_1,\ c_2) - G(c_1, c_2)| \leq K\delta_0. \tag{2.3}$$

Proof. By (2.1), for any $\varepsilon > 0$ there exists $\pi \in \Pi$ such that

$$g_i(c_1,\ c_2) + \varepsilon \geq E_\pi \left(G(c_1 + g^*\widetilde{\tau}_N,\, c_2 + \widetilde{D}_N) \middle| X_0 = i \right). \tag{2.4}$$

For simplicity, put $P(\cdot) = P_\pi(\cdot|X_0 = i)$, $E(\cdot) = E_\pi(\cdot|X_0 = i)$ and $H_n = (X_0,\, \Delta_0, \tau_1,\, X_1, \cdots, X_n)$. We have the following:

$$\begin{aligned}
&E\left[G(c_1 + g^*\widetilde{\tau}_N,\, c_2 + \widetilde{D}_N)|N = n\right] \\
&= E\left[E\left[G(c_1 + g^*\widetilde{\tau}_N,\, c_2 + \widetilde{D}_N)\middle|N = n, H_{n-1}\right]|N = n\right] \\
&\geq E\left[E\left[G(c_1 + g^*\widetilde{\tau}_{n-1},\, c_2 + \widetilde{D}_{n-1})|N = n, H_{n-1}\right]\middle|N = n\right] - K \\
&\qquad\qquad (\text{ from Assumption 3 }) \\
&= E\left[G(c_1 + g^*\widetilde{\tau}_{N-1},\, c_2 + \widetilde{D}_{N-1})\middle|N = n\right] - K \\
&\ \vdots \qquad\qquad (\text{ repeating the same discussion }) \\
&\geq G(c_1, c_2) - nK.
\end{aligned}$$

Thus, it follows that

$$\begin{aligned}
&\quad E\left[G(c_1 + g^*\widetilde{\tau}_N,\, c_2 + \widetilde{D}_N)\right] \\
&= \sum_{n=0}^{\infty} P(N = n) E\left[G(c_1 + g^*\widetilde{\tau}_N,\, c_2 + \widetilde{D}_N)\middle|N = n\right] \\
&\geq G(c_1, c_2) - K\sum_{n=0}^{\infty} nP(N = n) \\
&\geq G(c_1, c_2) - K\sum_{n=0}^{\infty} e(n) \\
&= G(c_1, c_2) - K\delta_0.
\end{aligned}$$

From (2.4), we find that $g_i(c_1,\ c_2) + \varepsilon \geq G(c_1, c_2) - K\delta_0$. As $\varepsilon \to 0$ in the above, we get $g_i(c_1,\ c_2) \geq G(c_1, c_2) - K\delta_0$. Starting from $g_i(c_1,\ c_2) \leq$

$E_\pi(G(c_1 + g^*\widetilde{\tau}_N,\ c_2 + \widetilde{D}_N)|X_0 = i)$ for a policy $\pi \in \Pi$, apply the same way as the above discussion. Then, we get $g_i(c_1,\ c_2) \le G(c_1,\ c_2) + K\delta_0$. $\square$

Lemma 2.2. *There exists a* $\overline{K} > 0$ *such that*

$$\int_0^\infty \Big| g_i(\overline{c_1},\ \overline{c_2}) - g_i(c_1,\ c_2) \Big| F_{ij}(dt|\, a) \le \overline{K} \tag{2.5}$$

where $\overline{c_1} = c_1 + g^*t, \overline{c_2} = c_2 + r(i,\ a) + d(i,\ a)t$ *for all* $c_1,\ c_2 \in \boldsymbol{R}_+,\ \ i,\ j \in S$ *and* $\ a \in A$.

Proof. We have that

$$\begin{aligned} |g_i(\overline{c_1},\ \overline{c_2}) - g_i(c_1,\ c_2)\,| \ &\le |g_i(\overline{c_1},\ \overline{c_2}) - G(\overline{c_1},\ \overline{c_2})\,| + |g_i(c_1,\ c_2) - G(c_1,\ c_2)\,| \\ &\quad + |G(\overline{c_1},\ \overline{c_2}) - G(c_1,\ c_2)\,|\,. \end{aligned}$$

So, from Lemma 2.1 and Assumption 3, the inequality (2.5) holds with $\overline{K} = K(2\delta_0 + 1)$. $\square$

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

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

where $\overline{c_1} = c_1 + g^*t, \overline{c_2} = c_2 + r(i,\ a) + d(i,\ a)t$ 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 our main results, which gives the optimality equation and characterizes the regret optimal policies.

Theorem 2.1. (i) *The set of optimal value functions* $g = (g_i : i \in J)$ *satisfies 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* $c_1,\ c_2 \in \boldsymbol{R}_+$.

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

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

for all $n \ge 0$ *and* H_n, *where* $\ A^*(c_1, c_2 : i) = \mathrm{argmin}_{a\in A} U\{g\}(c_1, c_2|i, a)$ *for* $c_1, c_2 \in \boldsymbol{R}_+$ *and* $i \in J$. *Then,* π^* *is regret optimal with a target* g^*.

Proof. For (i), for any $\varepsilon > 0$, $i,\, j \in S$, $a \in A$ and $t \in \boldsymbol{R}_+$, there exists a policy $\pi\{i,\, a,\, t,\, j\} = (\pi\{i,\, a,\, t,\, j\}_0,\, \pi\{i,\, a,\, t,\, j\}_1,\, \cdots\cdots)$ satisfying that

$$g_j(\overline{c_1},\, \overline{c_2}) + \varepsilon \geq E_{\pi\{i,\, a,\, t,\, j\}} \left(G(\overline{c_1} + g^* \widetilde{\tau}_N,\ \overline{c_2} + \widetilde{D}_N) \Big|\, X_0 = j \right) \tag{2.9}$$

where $\overline{c_1} = c_1 + g^* t, \overline{c_2} = c_2 + r(i,\, a) + d(i,\, a)t$. Here we define a policy $\pi' = (\pi'_0,\, \pi'_1, \cdots)$ by $\pi'_0(a|H_0) = 1$, $\pi'_n(\cdot|H_n) = \pi\{X_0, \Delta_0, \tau_1, X_1\}_{n-1}(\cdot|H'_{n-1})$ for $n \geq 1$, where $H'_{n-1} = (X_1,\, \Delta_1,\, \tau_2,\, X_2, \cdots, X_n)$ is shifted from $H_n = (X_0,\, \Delta_0,\ \tau_1, X_1, \cdots, X_n)$. Then, we have from (2.1), (2.9) and (2.6) that

$$\begin{aligned}
& g_i(c_1,\, c_2) \\
\leq\ & E_{\pi'} \left(G(c_1 + g^* \widetilde{\tau}_N, c_2 + \widetilde{D}_N) \Big| X_0 = i \right) \\
=\ & \sum_{j \in S} p_{ij}(a) \int_0^\infty E_{\pi\{i,a,t,j\}} \left(G(\overline{c_1} + g^* \widetilde{\tau}_N, \overline{c_2} + \widetilde{D}_N) \Big| X_0 = j \right) F_{ij}(dt|a) \\
\leq\ & \varepsilon + \sum_{j \in S} p_{ij}(a) \int_0^\infty g_j(\overline{c_1}, \overline{c_2}) F_{ij}(dt|a) \\
=\ & \varepsilon + U\{g\}(c_1,\, c_2|\, i,\, a).
\end{aligned}$$

Since $\varepsilon > 0$ and $a \in A$ are arbitrary, we get

$$g_i(c_1,\, c_2) \leq \min_{a \in A} U\{g\}(c_1,\, c_2|\, i,\, a). \tag{2.10}$$

On the other hand, for any $\varepsilon > 0$, there exists a $\pi = (\pi_0,\, \pi_1, \cdots) \in \Pi$ such that

$$\begin{aligned}
& g_i(c_1,\, c_2) + \varepsilon \\
\geq\ & E_\pi \left(G(c_1 + g^* \widetilde{\tau}_N, c_2 + \widetilde{D}_N) |\, X_0 = i \right) \\
=\ & \sum_{a,j} \pi_0(a|i) p_{ij}(a) \int_0^\infty E_{\pi\{i,a,t,j\}} \left(G(\overline{c_1} + g^* \widetilde{\tau}_N, \overline{c_2} + \widetilde{D}_N) \Big| X_0 = j \right) F_{ij}(dt|a) \\
\geq\ & \min\nolimits_{a \in A} U\{g\}(c_1,\, c_2|\, i,\, a).
\end{aligned}$$

A conditional policy $\pi\{i, a, t, j\} = (\pi\{i, a, t, j\}_k; k = 0, 1, 2, \cdots)$ means that $\pi\{i, a, t, j\}_n(\cdot|H_n) = \pi_{n+1}(\cdot|i, a, t,\, H_n)$ where $H_n = (X_0 = j,\, \Delta_0, \cdots, X_n)$ for $n \geq 0$. Combined with (2.10), this last inequailty shows that (2.7) holds.

For (ii), put $P(\cdot) = P_{\pi^*}(\cdot|X_0 = i)$ and $E(\cdot) = E_{\pi^*}(\cdot|X_0 = i)$ for simplicity. Then, we have from (2.7) that, for $n > 0$,

$$\begin{aligned}
& E \left(g_{X_{n+1}}(g^* \widetilde{\tau}_{n+1},\ \widetilde{D}_{n+1}) \mathbf{1}_{N > n} \Big|\, H_n,\, \Delta_n \right) \\
=\ & U\{g\}(g^* \widetilde{\tau}_n,\ \widetilde{D}_n|\, X_n,\, \Delta_n\,) \mathbf{1}_{N > n} \\
=\ & g_{X_n}(g^* \widetilde{\tau}_n,\ \widetilde{D}_n) \mathbf{1}_{N > n},
\end{aligned} \tag{2.11}$$

where $\mathbf{1}_A$ is the indicator of a set A. So, we get that

$$
\begin{aligned}
& E\left(g_{X_n}(g^*\widetilde{\tau}_n,\ \widetilde{D}_n)\mathbf{1}_{N>n}\right) \\
= \ & E\left(E(g_{X_n}(g^*\widetilde{\tau}_n,\ \widetilde{D}_n)\mathbf{1}_{N>n}|H_n,\, \Delta_n\,)\right) \\
= \ & E\left(E(g_{X_{n+1}}(g^*\widetilde{\tau}_{n+1},\ \widetilde{D}_{n+1})\mathbf{1}_{N>n}|H_n,\, \Delta_n\,)\right) \\
= \ & E\left(g_{X_{n+1}}(g^*\widetilde{\tau}_{n+1},\ \widetilde{D}_{n+1})\mathbf{1}_{N=n+1}\right) + E\left(g_{X_{n+1}}(g^*\widetilde{\tau}_{n+1},\ \widetilde{D}_{n+1})\mathbf{1}_{N>n+1}\right).
\end{aligned}
$$

Repeating the above discussion, we have that

$$
\begin{aligned}
& E\left(g_{X_n}(g^*\widetilde{\tau}_n, \widetilde{D}_n)\mathbf{1}_{N>n}\right) \\
= & \sum_{k=n+1}^{\ell} E\left(g_{X_k}(g^*\widetilde{\tau}_k, \widetilde{D}_k)\mathbf{1}_{N=k}\right) + E\left(g_{X_\ell}(g^*\widetilde{\tau}_\ell, \widetilde{D}_\ell)\mathbf{1}_{N>\ell}\right).
\end{aligned} \tag{2.12}
$$

Also, we have from Lemma 2.1, 2.2 and Assumption 3 that

$$
\begin{aligned}
E\left(g_{X_\ell}(g^*\widetilde{\tau}_\ell, \widetilde{D}_\ell)\mathbf{1}_{N>\ell}\right) &= P(N>\ell)E\left(g_{X_\ell}(g^*\widetilde{\tau}_\ell, \widetilde{D}_\ell)\middle| N>\ell\right) \\
&\geq P(N>\ell)\left\{E\left(g_{X_{\ell-1}}(g^*\widetilde{\tau}_{\ell-1}, \widetilde{D}_{\ell-1})\middle| N>\ell\right) - \overline{K}\right\} \\
&\geq P(N>\ell)\left\{E\left(g_{X_1}(g^*\widetilde{\tau}_1, \widetilde{D}_1)\middle| N>\ell\right) - (\ell-1)\overline{K}\right\} \\
&\geq P(N>\ell)\{G(0,0)+\delta_0 K - \ell\overline{K}\}.
\end{aligned}
$$

Since $P(N>\ell)\to 0$ and $\ell P(N>\ell)\to 0$ as $\ell\to\infty$, for any $\varepsilon>0$ there exists ℓ_0 such that $E(g_{X_\ell}(g^*\widetilde{\tau}_\ell,\ \widetilde{D}_\ell)\mathbf{1}_{N>\ell}) > -\varepsilon$ for all $\ell\geq\ell_0$. Also, since $g_{X_k}(g^*\widetilde{\tau}_k,\ \widetilde{D}_k)\mathbf{1}_{N=k} = G(g^*\widetilde{\tau}_k,\ \widetilde{D}_k)$, (2.12) implies that

$$
E(g_{X_n}(g^*\widetilde{\tau}_n,\ \widetilde{D}_n)\mathbf{1}_{N>n}) \geq \sum_{k=n+1}^{\ell} E\left(G(g^*\widetilde{\tau}_k,\ \widetilde{D}_k)\mathbf{1}_{N=k}\right) - \varepsilon. \tag{2.13}
$$

By the above with $n=0$, we get that

$$
g_i(0,\,0) \geq E\left(G(g^*\widetilde{\tau}_N,\ \widetilde{D}_N)\mathbf{1}_{N\leq\ell}\right) - \varepsilon \tag{2.14}
$$

for all $\ell\geq\ell_0$. As $\ell\to\infty$ and $\varepsilon\to 0$ in (2.14), it holds that $g_i(0,0)\geq E\left(G(g^*\widetilde{\tau}_N, \widetilde{D}_N)\right)$. Obviously, $g_i(0,0)\leq E\left(G(g^*\widetilde{\tau}_N, \widetilde{D}_N)\right)$, so that $g_i(0,0) = E\left(G(g^*\widetilde{\tau}_N, \widetilde{D}_N)\right)$, which shows that π^* is regret optimal. □

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 $\boldsymbol{C} = \{h = (h_i : i \in J) | \, h_i \in \mathcal{B}(\boldsymbol{R}_+ \times \boldsymbol{R}_+)$ for all $i \in J$ and h satisfies the statement of Lemma 2.2.*

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 $\overline{a} \in A$ such that

$$
\begin{aligned}
& \left| h_i(c_1,\, c_2) - h'_i(c_1,\, c_2) \right| \\
\le \;& \sum_{j \in J} p_{ij}(\overline{a}) \left| \int_0^\infty h_j(\overline{c_1}, \overline{c_2}) F_{ij}(dt|\,\overline{a}) - \int_0^\infty h'_j(\overline{c_1}, \overline{c_2}) F_{ij}(dt|\,\overline{a}) \right| \\
\le \;& \sum_{j \in J} p_{ij}(\overline{a}) \left(\left| h_j(c_1,\, c_2) - h'_j(c_1,\, c_2) \right| + 2\overline{K} \right).
\end{aligned}
\tag{2.15}
$$

Repeating the relation (2.15), we get that

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

So, if we put $\|h_i - h'_i\| = \sup_{c_1,\, c_2 \in R_+} \left| h_i(c_1,\, c_2) - h'_i(c_1,\, c_2) \right|$, then $\|h_i - h'_i\| \le 2\overline{K}\delta_0$, and from the first inequality in (2.15), we get

$$
\left\| h_i - h'_i \right\| \le \sum_{j \in J} p_{ij}(\overline{a}) \left\| h_j - h'_j \right\| \quad \text{for} \quad i \in J. \tag{2.16}
$$

Repeating (2.16) again, we obtain

$$
\left\| h - h' \right\| \le e(n) \left\| h - h' \right\| \quad \text{for all} \quad n \ge 1, \tag{2.17}
$$

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, it means $\left\| h - h' \right\| = 0$. 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

$$g_i(c_1, c_2) = \inf_{\pi\in\Pi} E_\pi(c_1 + g^*\widetilde{\tau}_N - c_2 - \widetilde{D}_N \mid X_0 = i) = c_1 - c_2 + g_i \tag{3.1}$$

for $i \in J$ and $c_1, c_2 \in \boldsymbol{R}_+$. 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 $i \in J = S \setminus J_0$ with some absorbing state J_0, where $R(i, a) = r(i, a) + d(i, a)\overline{\tau}(i, a)$ and $\overline{\tau}(i, a) = \sum_{j\in S} p_{ij}(a) \int_0^\infty tF_{ij}(dt|a)$ for $i \in J$ 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:

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

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

Then, (3.2) with $g_0 = 0$ is corresponding to the optimality equation for the average case. In fact, it holds (cf. [15],[17]]) 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: $G(x, y) = -e^{-\lambda(x-y)}$, $(\lambda > 0)$. If the target value g^* is sufficiently large such that $g^*t - r(i, a) - d(i, a)t \geq 0$ is satisfied for all $t \geq 0$, $i \in S$ and $a \in A$, Assumption 3 in Section 2 holds obviously. Let

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

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) becomes

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

where $R(i, a, j) = \displaystyle\int_0^\infty e^{-\lambda(g^*t - r(i,a) - d(i,a)t)} F_{ij}(dt|a)$ for $i \in J$, $j \in S, a \in A$. Applying Theorem 2.1, we get a regret optimal policy for the exponential regret-utility case.

Acknowledgements: The authors should express thanks to anonymous reviewer who give us useful comments and practical references. These help us to improve our earlier draft of paper.

References

[1] Bielecki, T., Hernandez-Hernadez, D. and Pliska,S.R. (1999): Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management, *Math. Meth. of Oper. Res.*, **50**, 167-188.

[2] Borkar, V.S. and Meyn, S.P. (2002): Risk-sensitive optimal control for Markov decision processes with monotone cost, *Math. Oper. Res.*, **27**, 192-209.

[3] Cavazos-Cadena, R. and Montos-De-Oca, R. (2003): The value iterarion algorithm in risk-sensitive average Markov decision chains with finite state space, *Math. Oper. Res.*, **28**, 752-776.

[4] Cavazos-Cadena, R. and Fernades-Gaucherand, E. (1999): Controlled Markov chains with risk-sensitive criteria: Average cost, optimality equations, and optimal solutions, *Math. Meth. Oper. Res.*, **49**, 299-324.

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

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

[7] Filar, J.A., Kallenberg,L.C.M. and Huey-Min Lee (1989) Variance-Penalized Marikov Decision Processes, *Math. of Oper. Res.* , **14**, 147-161.

[8] Filar,J.A., Krass,D. and Ross, K.W. (1995): Percentaile Performance Criteria For Limiting Average Markov Decision Processes, *IEEE AC*, **40**, 2-10.

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

[10] 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

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

[12] 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.

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

[14] Lin,Yuanlie, Filar,J.A. and Ke Liu (2002): Finite Horizon Portfolio Risk Models with Probability Criterion, In: Hou et al(eds) *Markov Processes and Controlled Markov Chains*, Kluwer Academic Publishers, 405-424.

[15] 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.

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

[17] Ross,S.M. (1970): Applied Probability Models with Optimization Applications, Holden-Day, San Francisco.

[18] White,D.J. (1988): Mean, variance and probabilistic criteria in finite Markov decision processes: A review, *J.Optim. Theory Appl.*, **56**, 1-29.

[19] White,D.J. (1993): Minimizing a threshold probability in discounted Markov decision processes, *J.Math. Anal. Appl.*, **56**, 1-29.

[20] Whittle,P. (1990): Risk-sensitive Optimal Control, Wiley, New York.

[21] Wu,Congbin and Lin,Yuanlie (1999): Minimizing Risk Models in Markov Decision Processes With Policies Depending on Target Values, *J. Math. Anal. Appl.*, **231**, 47-67.

[22] Yu,S.X., Yuanlie Lin and Pingfan Yan(1998): Optimization Models for the First Arrival Target Distribution Function in DIscrete Time, *J.Math.Anal.Appl.*, **225**, 193-223.