Stopped Decision Processes in conjunction with General Utility

Yoshinobu KADOTA
Dep. of Math., Wakayama Univ., Wakayama 640, Japan
yoshi-k@math.edu.wakayama-u.ac.jp
Masami KURANO
Dep. of Math., Chiba Univ., Chiba 263-8522, Japan
kurano@math.e.chiba-u.ac.jp
Masami YASUDA
Dep. of Math. & Inform., Chiba Univ., Chiba 263-8522, Japan
yasuda@math.s.chiba-u.ac.jp

Abstract

A combined model of Markov decision process and the stopping problem, called Stopped Decision Processes, with a general utility is considered in the paper. This model is important for the various applications including Bandit problems and piecewise deterministic Markov processes, etc. Our aim is to formulate the general utility-treatment of processes with a countable state space and give a functional characterization from points of seeking an optimal pair, that is, a policy and a stopping time. And the further results concerning an optimality equation of our model are given. If we restrict the problem as the choice of a policy for a given fixed stopping time, this results are consistent with our previous paper(1998). Especially, in the case of the exponential utility functions, the optimal pair can be derived concretely using the idea of the one-step look ahead policy. Also, a simple numerical example is given to illustrate the results.

Keywords: *Markov Decision Process, Optimal Stopping Problem, Stopped Decision Process, General Utility, Optimal Pair, Optimality Equation, Exponential Utility.*

1 Introduction and formulation

Markov decision processes are applied to various problems such as an economic dynamics, Bandit problem, queuing net works, etc. and many papers are published. See [16, 17, 19]. Also the optimal stopping problem is now discussed to the application of stock option markets so extensive studies are executed([2, 5]). In this paper a combined model of Markov decision process and the stopping problem, called Stopped Decision Processes, is considered in conjunction with general utility. The general utility-treatment of stopped decision processes with a countable state space seem not to be considered yet.

There are some papers[4, 9] which combined with two notions recently. However Furukawa and Iwamoto[7] had discussed for reward system of the additive type and Furukawa[8] has reformulated the stopped decision model in the fashion of gambling theory. Hordijk[10] also had considered this model from a standpoint of potential theory.

Our previous paper[14] has already considered the optimization problem of the expected utility for the total discounted reward random variable accumulated until the stopping time. We derived an optimality equation of the general utility case, by which an optimal pair, that is, a policy and a stopping time has been characterized, however, the concrete method of seeking an optimal pair is not discussed there.

The objective of this paper is to give a functional characterization from points of view of seeking an optimal pair. And we give further results concerning an optimality equation of this model, which is useful in seeking an optimal pair. Also, for the case of the exponential utility function (cf. [6, 15]), the optimal pair can be obtained concretely using the idea of the one-step look ahead policy(cf. [17]). By using an idea of which is appearing in Chung and Sobel[3], Sobel[18] and White[20], the optimality equation is described by the class of distribution functions of the present value.

In the remainder of this section, we will formulate the problem to be examined and an optimal pair of a policy and a stopping time is defined. Firstly a well-known formulation of a standard Markov decision processes (MDPs) is concerned with

$$(S, \{A(i)\}_{i\in S}, q, r),$$

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

Throughout this paper we assume as follows:

(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 $r(i, \cdot, j)$ are continuous on $A(i)$, and

(iii) $r(\cdot, \cdot, \cdot)$ is uniformly bounded.

A sample space of the decision processes 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 time t of the process $(t \geq 0)$. A policy $\pi = (\pi_0, \pi_1, \cdots)$ is a sequence of conditional probabilities π_t such that $\pi_t(A(i_t) \mid i_0, a_0, \cdots, i_t) = 1$ for all histories $(i_0, a_0, \cdots, i_t) \in (S \times A)^t \times S$. The set of policies is denoted by Π. Let $H_t = (X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t)$ for $t \geq 0$. We assume for each $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$,

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

which is independent of the last history H_{t-1} and the action Δ_{t-1} 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 measure $\nu \in \mathcal{P}(S)$ and a policy $\pi \in \Pi$ determine the probability measure $P_\nu^\pi \in \mathcal{P}(\Omega)$ by a usual way. *A total present value* until time t is defined by

$$\mathcal{B}(t) := \sum_{k=0}^{t} r(X_{k-1}, \Delta_{k-1}, X_k) \quad (t \geq 0), \tag{1.1}$$

where X_{-1}, Δ_{-1} are fictitious variables and set $r(X_{-1}, \Delta_{-1}, X_0) = 0$. Note that the total present value is a random variable defined on the probability space (Ω, P_ν^π) for each $\nu \in \mathcal{P}(S)$ and $\pi \in \Pi$.

In order to induce a utility function of the problem in the reward structure, let us consider g as a non-decreasing continuous function on the real space $\mathcal{R}$. And we call a random variable $\sigma : \Omega \to \{0, 1, 2, \cdots\}$ *a stopping time w.r.t.* (ν, π) if the following conditions are satisfied:

(i) For each $t \geq 0$, $\{\sigma = t\} \in \mathcal{F}(H_t)$,

(ii) $P_\nu^\pi(\sigma < \infty) = 1$ and

(iii) $E_\nu^\pi[g^-(\mathcal{B}(\sigma))] < \infty$,

where $\mathcal{F}(H_t)$ is the σ-algebra induced by H_t and $g^{\pm}(x) = \max\{\pm g(x), 0\}$. The class of this stopping times will be denoted by $\Sigma_{(\nu,\pi)}$ hereafter. For any $\nu \in \mathcal{P}(S)$, let

$$\mathcal{A}_\nu := \{(\pi, \sigma) \mid \sigma \in \Sigma_{(\nu,\pi)}, \pi \in \Pi\}.$$

Our problem is to maximize the expected utility $E_\nu^\pi[g(\mathcal{B}(\sigma))]$ over all $(\pi, \sigma) \in \mathcal{A}_\nu$ for a fixed $\nu \in P(S)$. Therefore the pair of a policy and a stopping time, $(\pi^*, \sigma^*) \in \mathcal{A}_\nu$, is called (ν, g)-*optimal* or simply *optimal pair* if

$$E_\nu^{\pi^*}[g(\mathcal{B}(\sigma^*))] \geq E_\nu^\pi[g(\mathcal{B}(\sigma))] \quad \text{for all } (\pi, \sigma) \in \mathcal{A}_\nu. \tag{1.2}$$

In Section 2, we will give a characterization of the optimal policy. In the case that the stopping time is fixed, its results are applied to obtain an optimal pair in the sequel section. In Section 3, we extend the results obtained in [14] for the discount reward to the general case. An optimality equation for the stopped decision process is derived and the optimal pair is characterized. The proofs are analogous to those in [14], so that several proofs in the theorems will be omitted. The exponential utility case is treated in Section 4, where the optimal pair is sought by using the idea of the one-step look ahead (OLA) stopping time.

2 MDPs with the stopping region

For a subset K of the state space S, let us consider a stopping time of the first hitting time for the region, that is,

$$\sigma_K := \text{ the first time } t \geq 0 \text{ such that } X_t \in K.$$

Henceforth we assume that σ_K is a stopping time w.r.t. any $(\nu, \pi) \in \mathcal{P}(S) \times \Pi$. We say that a policy $\pi^* \in \Pi$ is *(ν, g)-optimal w.r.t.* σ_K if

$$E_\nu^{\pi^*}[g(\mathcal{B}(\sigma_K))] \geq E_\nu^{\pi}[g(\mathcal{B}(\sigma_K))] \quad \text{for all } \pi \in \Pi.$$

When π^* is (ν, g)-optimal for all $\nu \in \mathcal{P}(S)$, it is simply called *g-optimal w.r.t.* σ_K. In order to analyze the above problem, it is convenient for us to rewrite $E_\nu^\pi[g(\mathcal{B}(\sigma_K))]$ by using the distribution function of $\mathcal{B}(\sigma_K)$ corresponding to P_ν^π. Suppressing K in the notation, let for $\nu \in \mathcal{P}(S)$ and $\pi \in \Pi$,

$$F_\nu^\pi(z) := P_\nu^\pi(\mathcal{B}(\sigma_K) \leq z) \quad \text{and}$$

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

Then, it is obvious that $E_\nu^\pi[g(\mathcal{B}(\sigma_K))] = \displaystyle\int g(z)\, F_\nu^\pi(dz)$ holds.

For any $\pi \in \Pi$, the following integral is well-defined and hence the map $v_\pi : \mathcal{R} \times \mathcal{P}(S) \to \mathcal{R}$ means a value function associated with a fortune d and an initial measure ν :

$$v_\pi(d, \nu) := \int g_d(z)\, F_\nu^\pi(dz) := \int g(d+z)\, F_\nu^\pi(dz)$$

where $g_d(z) := g(d+z)$. We note that $v_\pi(0, \nu) = E_\nu^\pi[g(\mathcal{B}(\sigma_K))]$ and $v_\pi(d, i) = g(d)$ if $i \in K$, where the initial measure $\nu \in \mathcal{P}(S)$ is simply denoted by i whenever if the measure is degenerate at $\{i\}$.

The value function for our model can be denoted by

$$v(d, \nu) := \sup_{\pi \in \Pi} v_\pi(d, \nu), \tag{2.1}$$

which is depending on a *present fortune* $d \in \mathcal{R}$ and a *state distribution* $\nu \in \mathcal{P}(S)$.

In the following lemma, it is shown that the supremum in (2.1) can be attainable.

Lemma 2.1. *For any* $\nu \in \mathcal{P}(S)$, $v(d, \nu) = \max_{F \in \Phi(\nu)} \displaystyle\int g_d(z)\, F(dz)$ holds, *and, for each* $\nu \in \mathcal{P}(S)$, *there exists* (ν, g)*-optimal policy w.r.t.* σ_K.

Proof. For each $\nu \in \mathcal{P}(S)$, the set $\{P_\nu^\pi(\cdot) \in \mathcal{P}(\Omega) \mid \pi \in \Pi\}$ is known to be compact in the weak topology(c.f. Borker[1]). Since the map $\mathcal{B}(\sigma_K) : \Omega \to \mathcal{R}$ is continuous, $\Phi(\nu)$ is weak-compact. Thus, from the assumption of the continuity of $g_d(z)$, it follows that

$$v(d, \nu) = \sup_{F \in \Phi(\nu)} \int g_d(z)\, F(dz) = \int g_d(z)\, F^*(dz)$$

for some $F^* \in \Phi(\nu)$. This proves the first part of the results. For $F^* \in \Phi(\nu)$ with $v(0, \nu) = \displaystyle\int g(z)\, F^*(dz)$, the policy π^* corresponding to F^* is clearly (ν, g)-optimal w.r.t. σ_K, as required. $\square$

Lemma 2.2. *For each $t \geq 0$, $d \in \mathcal{R}$ and $\pi \in \Pi$,*

$$
\begin{aligned}
&E_\nu^\pi[g_d(\mathcal{B}(\sigma_K)) \mid \sigma_K > t] \\
&\leq E_\nu^\pi\Big[\max_{a \in A(X_t)} \sum_{j \in S} q_{X_t,j}(a)\, v(\tilde{d}, j) \mid \sigma_K > t\Big],
\end{aligned}
\tag{2.2}
$$

where $\tilde{d} := d + \mathcal{B}(t-1) + r(X_t, a, j)$.

Proof. For simplicity, denote E_ν^π 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 $t \geq 1$. The Markov property of the transition law yields that

$$
\begin{aligned}
& E[g_d(\mathcal{B}(\sigma_K)) \mid \sigma_K > t] \\
= \;& E[E[g_d\big(\mathcal{B}(t-1) + r(X_t, \Delta_t, X_{t+1}) + \mathcal{B}(\sigma_K)(\theta_t(\omega))\big) \mid H_t] \mid \sigma_K > t] \\
\leq \;& E[v(d + \mathcal{B}(t-1) + r(X_t, \Delta_t, X_{t+1})) \mid \sigma_K > t] \\
\leq \;& \{\text{the right-hand of the inequality (2.2)}\},
\end{aligned}
$$

which completes the proof. □

For any $d \in \mathcal{R}$ and $i \notin K$, let

$$A(d,i) := \arg\max_{a \in A(i)} \sum_{j \in S} q_{ij}(a)\, v(d + r(i,a,j),\, j).$$

The value function $v(d,i)$ is shown to satisfy the optimality equation in the following theorem.

Theorem 2.1. *The value function $v(d,i); d \in \mathcal{R}, i \in S$ satisfies the following equation.*

$$
v(d,i) = \begin{cases} \max_{a \in A(i)} \sum_{j \in S} q_{ij}(a)\, v(d + r(i,a,j),\, j) & \text{for } \; i \notin K, \\ d & \text{for } \; i \in K. \end{cases}
\tag{2.3}
$$

Proof. For any $f \in F$ such that $f(i) \in A(d,i)$ for $i \notin K$ and $f(i) = a$(arbitrary)$\in A(i)$ for $i \in K$, let $\pi^{(j)}$ be a policy corresponding to $F_j^* \in \Phi(j)$ satisfying

$$v(d + r(i, f(i), j),\, j) = \int g_d\big(r(i, f(i), j) + z\big)\, F_j^*(dz),$$

whose existence is guaranteed by Lemma 2.1. Let π be the policy that choose the action Δ_0 at time 0 according to f and use policy $\pi^{(j)}$ from time 1 when $X_1 = j$. Then, clearly it holds

$$
\begin{aligned}
E_i^\pi[g_d(\mathcal{B}(\sigma_K))] &= \textstyle\sum_{j \in S} q_{ij}(f(i))\, v(d + r(i, f(i), j),\, j) \\
&= \max_{a \in A(i)} \textstyle\sum_{j \in S} q_{ij}(a)\, v(d + r(i, f(i), j),\, j).
\end{aligned}
$$

Together with (2.2), the above derives (2.3), as required. □

In order to discuss the uniqueness of solutions of (2.3), we need the following assumption.

Assumption A. $E_\nu^\pi[\mid g_d(\mathcal{B}(\sigma_K)) \mid] < \infty$ for any $\nu \in \mathcal{P}(S)$, $\pi \in \Pi$ and $d \in \mathcal{R}$.

Theorem 2.2. *Suppose that Assumption A holds.*

(i) *It holds that, for any $\nu \in \mathcal{P}(S)$, $\pi \in \Pi$ and $d \in \mathcal{R}$,*

$$\lim_{t\to\infty} E_\nu^\pi[v(d+\mathcal{B}(t),\, X_t)\mathbf{1}_{\{\sigma_K>t\}}] = 0 \tag{2.4}$$

where $\mathbf{1}_A$ is an indicator function of a set A.

(ii) *The map $v : \mathcal{R} \times S \longrightarrow \mathcal{R}$ satisfying (2.4) is uniquely determined by (2.3).*

Proof. Let $\pi \in \Pi$. Let $\pi\{X_t\} \in \Pi$ be such that $v(d+\mathcal{B}(t),\, X_t) = v_{\pi\{X_t\}}(d+\mathcal{B}(t),\, X_t)$. Note that $\pi\{X_t\}$ is depending on $d+\mathcal{B}(t)$ and X_t and its existence is guaranteed by Lemma 2.1. We denote by $\pi^{(t)} \in \Pi$ the policy that uses π until time t and uses $\pi\{X_t\}$ from time t. Then,

$$\begin{aligned} & E_\nu^{\pi^{(t)}}[g_d(\mathcal{B}(\sigma_K))] \\ = \; & E_\nu^\pi[g_d(\mathcal{B}(\sigma_K))\mathbf{1}_{\{\sigma_K\le t\}}] + E_\nu^\pi[v(\sigma+\mathcal{B}(t),\, X_t)\mathbf{1}_{\{\sigma_K>t\}}]. \end{aligned} \tag{2.5}$$

Under Assumption A,

$$\lim_{t\to\infty} E_\nu^{\pi^{(t)}}[g_d(\mathcal{B}(\sigma_K))] = E_\nu^\pi[g_d(\mathcal{B}(\sigma_K))].$$

So as $t \to \infty$ in (2.5), we get (i). The proof of (ii) is not particularly difficult, but tedious. We omit the proof. □

The following results can be proved similarly as that of Theorem 3.3 in [12] so it is omitted.

Theorem 2.3. *Let $\pi^* = (\pi_0^*, \pi_1^*, \cdots)$ be any policy satisfying that for all $t \ge 0$*

$$\pi_t^*(A(\mathcal{B}(t),\, X_t) \mid H_t) = 1 \text{ on } \{\sigma_K > t\}.$$

Then π^* is g-optimal w.r.t. σ_K.

3 Optimal pairs

In this section we derive the optimality equation for the stopped decision model, by which an optimal pair is characterized. Throughout this section, we assume that the following conditions are satisfied.

Condition 1. The utility function g is differentiable and its derivative is weakly bounded. That is, for any compact subset D of $\mathcal{R}$, there exists a constant L_D such that

$$\mid g'(x) \mid \le L_D \quad \text{for all } x \in D.$$

Condition 2. The expectation of the utility function is upper bounded:

$$E_\nu^\pi[\sup_{t\geq 0} g^+(\mathcal{B}(t))] < \infty$$

for all $\nu \in \mathcal{P}(S)$ and $\pi \in \Pi$.

For simplicity of the notations, let

$$\widehat{\Phi}(\nu) := \{ F_\nu^{(\pi,\sigma)} \mid (\pi,\sigma) \in \mathcal{A}_\nu \},$$

where $F_\nu^{(\pi,\sigma)}(x) = P_\nu^\pi(\mathcal{B}(\sigma) \leq x)$ for $(\pi,\sigma) \in \mathcal{A}_\nu$. In order to describe an optimality equation in the sequel, let us define

$$U\{g\}(d,i,a,j) := \sup_{F\in\widehat{\Phi}(j)} \int g_d(r(i,a,j)+z)\, F(dz) \tag{3.1}$$

and

$$U_g(d,i) := \max_{a\in A(i)} \sum_{j\in S} q_{ij}(a)\, U\{g\}(d,i,a,j) \tag{3.2}$$

for each $d \in \mathcal{R}$, $i,j \in S$ and $a \in A(i)$. It is easily proved under Condition 1 that the maximum in (3.2) is attainable. For $\nu \in \mathcal{P}(S)$ and $n \geq 1$, let

$$\mathcal{A}_\nu^n := \{ (\pi, n\vee\sigma) \mid (\pi,\sigma) \in \mathcal{A}_\nu \}$$

where $a \vee b = \max\{a,b\}$ for $a,b \in \mathcal{R}$. Define a conditional maximum by

$$\gamma_n^\nu := \operatorname{esssup}_{(\pi,\sigma)\in\mathcal{A}_\nu^n} E_\nu^\pi[g(\mathcal{B}(\sigma)) \mid \mathcal{F}_n] \quad (n \geq 0),$$

where $\mathcal{F}_n = \mathcal{F}(H_n)$. Henceforth, for simplicity we write esssup by sup and suppress ν in γ_n^ν if not specified otherwise. The recursive relation concerning $\{\gamma_n\}$ is described in the following.

Lemma 3.1. ([14]) *For each $n \geq 0$, it holds*

(i) $\gamma_n = \max\{ g(\mathcal{B}(n)),\ \sup_{\pi\in\Pi} E_\nu^\pi[\gamma_{n+1} \mid \mathcal{F}_n]\}$

(ii) $\sup_{\pi\in\Pi} E_\nu^\pi[\gamma_{n+1}|\mathcal{F}_n] = U_g(\mathcal{B}(n), X_n)$.

In order to obtain an optimal pair, it is convenient to introduce the following notations:

$$\overline{\boldsymbol{R}} := \{ (d,i) \in \mathcal{R}\times S \mid g(d) \geq U_g(d,i) \} \quad \text{and}$$

$$A^*(d,i) := \arg\max_{a\in A(i)} \sum_{j\in S} q_{ij}(a)\, U\{g\}(d,i,a,j)$$

for each $d \in \mathcal{R}$ and $i \in S$.

Let

$$\sigma^* = \{\text{the first time } t \geq 0 \text{ such that } (\mathcal{B}_t, X_t) \in \overline{\boldsymbol{R}}\} \tag{3.3}$$

and $\pi^* = (\pi_0^*, \pi_1^*, \cdots)$ be any policy satisfying

$$P_\nu^{\pi^*}(\Delta_t \in A^*(\mathcal{B}(t), X_t)) = 1 \quad \text{for all} \quad t \geq 0. \tag{3.4}$$

The following lemma is given in [14].

Lemma 3.2. ([14]) *Let $\sigma^*(n) := \min\{\sigma^*, n\} \quad (n \geq 0)$. Then the process $\{(\gamma_{\sigma^*(n)}, \mathcal{F}_n); n \geq 0\}$ is a martingale.*

Here, we can state the main theorem.

Theorem 3.1. Under Condition 1 and 2, we have the following results:

(i) *If $P_\nu^{\pi^*}(\sigma^* < \infty) = 1$, then the pair (π^*, σ^*) is g-optimal.*

(ii) *If $g(\mathcal{B}(n)) \longrightarrow -\infty$ (as $n \to \infty$) $P_\nu^{\pi^*}$-a.s. then $P_\nu^{\pi^*}(\sigma^* < \infty) = 1$.*

Proof. From Lemma 3.2, $E_\nu^{\pi^*}[\gamma_0] = E_\nu^{\pi^*}[\gamma_{\sigma^*(n)}]$ for all $n \geq 1$. Now, as $n \to \infty$ in the above, we get

$$\begin{aligned} & E_\nu^{\pi^*}[\gamma_{\sigma^*}\mathbf{1}_{\{\sigma^*<\infty\}}] + E_\nu^{\pi^*}[\underline{\lim}_{n\to\infty}\gamma_n\mathbf{1}_{\{\sigma^*=\infty\}}] \\ \leq\ & E_\nu^{\pi^*}[\gamma_0] \\ \leq\ & E_\nu^{\pi^*}[\gamma_{\sigma^*}\mathbf{1}_{\{\sigma^*<\infty\}}] + E_\nu^{\pi^*}[\overline{\lim}_{n\to\infty}\gamma_n\mathbf{1}_{\{\sigma^*=\infty\}}]. \end{aligned} \tag{3.5}$$

If $P_\nu^{\pi^*}(\sigma^* < \infty) = 1$, $E_\nu^{\pi^*}[\gamma_0] = E_\nu^{\pi^*}[\gamma_{\sigma^*}]$. On the other hand, by the definition of σ^*, $\gamma_{\sigma^*} = g(\mathcal{B}(\sigma^*))$, which implies $E_\nu^{\pi^*}[\gamma_0] = E_\nu^{\pi^*}[g(\mathcal{B}(\sigma^*))]$. Since $E_\nu^{\pi}[g(\mathcal{B}(\sigma))] \leq E_\nu^{\pi}[\gamma_0] = E_\nu^{\pi^*}[\gamma_0]$, it holds $E_\nu^{\pi}[g(\mathcal{B}(\sigma))] \leq E_\nu^{\pi^*}[g(\mathcal{B}(\sigma^*))]$ for all $(\pi, \sigma) \in \mathcal{A}_\nu$. This shows that the pair (π^*, σ^*) is g-optimal. Thus the assertion (i) has proved. The assertion (ii) follows from (3.5). □

4 Exponential utility functions

In this section, we consider the case of the exponential utility function

$$g_\lambda(x) := \operatorname{sign}(-\lambda)\exp(-\lambda x) \tag{4.1}$$

for a constant $\lambda \neq 0$ and try to give the concrete characterization of the optimal pair by the OLA-stopping time (refer to Ross[17] and Kadota et al[13]).

Let

$$\eta_\lambda(i, a) := \sum_{j \in S} q_{ij}(a) \exp(-\lambda\, r(i, a, j)).$$

We need the following two conditions.

Condition A. For each $a \in A$, $\eta_\lambda(i, a)$ is non-decreasing in $i \in S$ if $\lambda > 0$, and alternatively, $\eta_\lambda(i, a)$ is non-increasing in $i \in S$ if $\lambda < 0$.

Condition B. For each $a \in A$, $q_{ij}(a) = 0$, if $i > j$ and $q_{ii}(a) < 1$.

We note that Condition B is satisfied for Markov deteriorating system. Let

$$K_\lambda := \{(d,i) \in \mathcal{R} \times S \mid v_\lambda(d,i) - g_\lambda(d) \leq 0\},$$

where $v_\lambda(d,i) = \max_{a \in A(i)} \sum_{j \in S} q_{ij}(a)\, g_\lambda(d + r(i,a,j))$. Then, the K_λ is characterized by the following.

Lemma 4.1. *Under Condition A, for each λ ($\lambda \neq 0$), there exists an integer $i_\lambda \in S$ such that $K_\lambda = \mathcal{R} \times \{i \in S \mid i \geq i_\lambda\}$.*

Proof. We observe that $v_\lambda(d,i) - g_\lambda(d) = e^{-\lambda d}(1 - \min_{a \in A(i)} \eta_\lambda(i,a))$ if $\lambda > 0$, $= e^{-\lambda d}(\max_{a \in A(i)} \eta_\lambda(i,a) - 1)$ if $\lambda < 0$. So that, if $\lambda > 0$, $v_\lambda(d,i) - g_\lambda(d) \leq 0$ means $\min_{a \in A} \eta_\lambda(i,a) \geq 1$. Observing that $\min_{a \in A} \eta_\lambda(i,a)$ is non-decreasing in $i \in S$ from Condition A, there exists an i_λ such that $v_\lambda(d,i) - g_\lambda(d) \leq 0$ if and only if $i \geq i_\lambda$. Similarly the case of $\lambda < 0$ is proved, as required. □

Lemma 4.2. *Under Condition A and B, the following holds:*

(i) *If $i \geq i_\lambda$, then, for all $a \in A(i)$ and $d \in \mathcal{R}$,*

$$\sum_{j \in S} q_{ij}(a)\, U\{g_\lambda\}(d,i,a,j) = \sum_{j \in S} q_{ij}(a)\, g_\lambda(d + r(i,a,j)) \tag{4.2}$$

(ii) *If $i < i_\lambda$, then $g_\lambda(d) < U_{g_\lambda}(d,i)$* for all $d \in \mathcal{R}$.

Proof. Let $i \geq i_\lambda$, $a \in A(i)$, $d \in \mathcal{R}$ and $j \in S$ with $q_{ij}(a) > 0$. Then, for any $F \in \widehat{\Phi}(j)$, from Condition B and Lemma 4.1 we see that $\{ g_\lambda(d + r(i,a,j) + \mathcal{B}(n)), \mathcal{F}_n, n = 0,1,2,\cdots\}$ is a super-martingale with respect to $\mathcal{P}_j^\pi$, where (π, σ) is a pair corresponding to F and $\mathcal{F}_n = \mathcal{F}(H_n)$. Applying Theorem 2.2 of Chow, Robbins and Siegmund[2], we get

$$\begin{aligned} E_i^\pi[g_\lambda(d + r(i,a,j) + \mathcal{B}(\sigma))] &= \int g_\lambda(d + r(i,a,j) + z)\, F(dz) \\ &\leq g_\lambda(d + r(i,a,j)), \end{aligned}$$

which means $U\{g_\lambda\}(d,i,a,j) \leq g_\lambda(d + r(i,a,j))$. Thus (4.2) follows. For (ii), let $i < i_\lambda$. Then, for any $d \in \mathcal{R}$, since $(d,i) \notin K_\lambda$ there exists $a_1 \in A(i)$ such that $g(d) < \sum_{j \in S} q_{ij}(a_1)\, g_\lambda(d + r(i,a_1,j))$. Thus, by the definition, clearly $g(d) < U_{g_\lambda}(d,i)$, as required. □

From Lemma 4.2, we find that the optimal stopping time σ_λ^* defined by (3.3) in Section 3 becomes

$$\sigma_\lambda^* := \text{ the first time } t \geq 0 \text{ with } X_t \in K_\lambda,$$

which is σ_{K_λ} with the stopping region K_λ and discussed in Section 2. Thus, to seek the optimal policy π^*, we can apply the results in Section 2.

Let

$$v_\lambda^*(i) := \mathrm{Opt}_{\pi\in\Pi} E_i^\pi[\exp\{-\lambda \mathcal{B}(\sigma_\lambda^*)\}]$$

where "Opt" means "Maximum" if $\lambda < 0$ and "Minimum" if $\lambda > 0$. Then, by Theorem 2.1 we have :
(4.3)

$$v_\lambda^*(i) = \begin{cases} \mathrm{Opt}_{a\in A(i)} \quad \Big[\sum_{i\le j\le i_\lambda} v_\lambda^*(j)\, q_{ij}(a)\, \exp\{-\lambda r(i,a,j)\} \\ \qquad\qquad + \sum_{j\ge i_\lambda} q_{ij}(a)\, \exp\{-\lambda r(i,a,j)\}\Big], & \text{for } i < i_\lambda, \\ 1, & \text{for } i \ge i_\lambda. \end{cases}$$

Let, for i $(1 \le i < i_\lambda)$,

$$A^*(i) := \{a \in A(i) \mid a \text{ realizes the opt on the RHS of (4.3)}\}$$

Then, the optimal pair $(\pi_\lambda^*, \sigma_\lambda^*)$ under exponential utility is given in the following theorem.

Theorem 4.1. *Let $\sigma_\lambda^* =$ the first t such that $X_t \ge i_\lambda$ and $\pi_\lambda^* = (\pi_0^*, \pi_1^*, \cdots)$ be such that $\pi_t^*\{A^*(X_t) \mid H_t\} = 1$ for $1 \le X_t < i_\lambda$. Then, the pair $(\pi_\lambda^*, \sigma_\lambda^*)$ is g-optimal.*

Proof. We can check that $\overline{\boldsymbol{R}}$ and $A^*(d,i)$ in Section 3 is equal to $\overline{\boldsymbol{R}}_\lambda$ and $A^*(i)$ respectively. Thus, from Theorem 3.1, Theorem 4.1 follows. □

Here we give a numerical example to illustrate the theoretical results. Let a countable state space $S = \{1, 2, 3, \cdots\}$, a action space of a closed interval $A = [1,\ 2]$ and

$$q_{ij}(a) = \begin{cases} \dfrac{(a/i)^{j-i}}{(j-i)!}\exp\{-a/i\}, & j \ge i \\ 0 & j < i, \end{cases}$$

for $i, j \in S$ and $a \in A$.

For an inspection cost $c > 0$, let $r(i,a,j) = a/i - c \quad (i, j \in S,\ a \in A)$. Then, $\eta_\lambda(i,a,j) = \exp\{-\lambda(a/i - c)\}$, which satisfies Condition A. Simple calculations yield the integer i_λ in Lemma 4.1 is given as $i_\lambda = \lceil 2/c \rceil + 1$, which is independent of λ, where $\lceil x \rceil$ is the smallest integer greater than or equal to x. Also, by (4.3) we find $A^*(i) = \{2\}$, so that the optimal policy $\pi_\lambda^* = 2^\infty$.

As another example, let $r(i,a,j) = (a/i)\,|j-i| - c$. Then, $\eta_\lambda(i,a) = \exp\{\lambda c + (a/i)\,(\exp(-\lambda a/i) - 1)\}$, which satisfies Condition A. The numerical value of each integer i_λ is given in Table 1.

Observing Table 1, we know that a risk-averse decision maker $(\lambda > 0)$ has a tendency to stop earlier than a risk-seeking one$(\lambda < 0)$.

	λ	-2.5	-2	-1.5	-1	-0.5		0.5	1	1.5	2	2.5
i_λ	c=0.1	8	8	8	7	7		7	6	6	6	6
	c=0.01	22	21	21	21	21		20	20	20	19	19

Table 1: The value of i_λ for $c = 0.1$ and $0.01(\lambda \neq 0)$.

References

[1] V. S. Borkar, *Topics in Controlled Markov Chains*, Longman Scientific Technical, 1991.

[2] Y. S. Chow, H. Robbins and D. Siegmund, *The Theory of Optimal Stopping : Great Expectations*, Houghton Mifflin Company, 1971.

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

[4] M.A.H.Dempster and J.J.Ye, Impulse control of piecewise deterministic Markov process, *Ann. Appl. Prob.*, **5** (1995) 399–423.

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

[6] P.C. Fishburn, *Utility Theory for Decision Making*, John Wiley & Sons, New York, 1970.

[7] N. Furukawa and S. Iwamoto, Stopped decision processes on complete separable metric spaces, *J. Math. Anal. Appl.*, **31** (1970) 615–658.

[8] N.Furukawa, Functional equations and Markov potential theory in stopped decision processes, *Mem. Fac. Sci, Kyushu Univ. Ser. A*, **29** (1975) 329–347.

[9] K.D.Glazebrook, Indices for families for competitive Markov decision processes with influence, *Ann. Appl. Prob.*, **3** (1993) 1013–1032.

[10] A. Hordijk, Dynamic Programming and Potential Theory, *Math. Centre Tracts No.* **51**, Mathematisch Centrum, Amsterdam, 1974.

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

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

[13] Y. Kadota, M. Kurano and M. Yasuda, Utility-Optimal Stopping in a denumerable Markov chain, *Bull. Inform. Cybernat.*, **28** (1995) 15–21.

[14] Y. Kadota, M. Kurano and M. Yasuda, On the general utility of discounted Markov decision processes. To appear in *International Transactions in Operational Reseach*, **4** (1998).

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

[16] M.L.Puterman, *Markov Decision Processes, Discrete Stochastic Dynamic Programming*, John Wiley & Sons, 1994.

[17] S.M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, 1970.

[18] M.J. Sobel, The variance of discounted Markov decision processes, *J. Appl. Prob.*, **19** (1982) 794–802.

[19] R.Weber, On the Gittens index for multiarmed bandits, *Ann. Appl. Prob.*, **2** (1992) 1024–1033.

[20] D.J. White, Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.*, **173** (1993) 634–646.