On the General Utility of Discounted Markov Decision Processes

(IFORS96 Pap.No.115)

Y.KADOTA, M.KURANO and M.YASUDA

Abstract

We will discuss an expected utility of rewards which are generated by Markov decision processes. This is applied to the optimal stopping problem with a utility treatment. Also a combined model of the decision processes and the stopping problem, called as a stopped Markov decision, is considered under the utility.

1 Introduction

Recently Stoky and Lucas[20] have applied Markov decision processes to analyze the economic dynamics. It has many attractive properties as Fishburn[5] and Pratt[16] etc. discuss by using the utility treatment. Several authors analyzed MDP's with exponential utility functions. Howard and Matheson[8] studied the case of finite states and actions in the finite horizon planning. They gave the policy improvement to find the policy that maximizes the time-average equivalent returns of MDP's. Chung and Sobel[3] 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[15] and Denardo and Rothblum[4] dealt with the problem from the other points of view.

In this paper, a utility optimization of Markov decision processes (MDP's) with countable state and compact action spaces is considered. Our aim is to construct a new combined model of the decision processes and the stopping problem, called as a stopped Markov decision, under the utility. In Section 2 some preliminaries are prepared to formulate Markov decision processes when the utility function is general. An optimality equation for the general utility and a necessary and sufficient condition of being optimal under the utility are derived. In Section 3 an optimal stopping problem with general utility is discussed. Inheriting the OLA policy from a usual stopping problem, the

validity of the policy is shown under suitable conditions. A property of the stopping time is characterized by the case of risk-averse or risk-seeking. These results are summarized from our previous papers, Kadota et al.[10,11].

Section 4 treats the main problem of the paper. On the basis of these previous arguments, the utility optimizations are induced by both of decision processes and stopping problems. We could extend the problem in the framework of general utility of discounted Markov decision processes.

2 A formulation of MDP with general utility functions

We consider standard Markov decision processes specified by

$$(S, \{A(i)\}_{i\in S}, q, r), \tag{2.1}$$

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 remain 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 $r(i, \cdot, j)$ are continuous on $A(i)$.*
(iii) *The function r is uniformly bounded, i.e., $0 \le r(i, a, j) \le M$ for all $i, j \in S$ and $a \in A(i)$.*

A sample space is the product space $\Omega = (S \times A)^\infty$ such that the projections X_t, Δ_t on the t-th factors S, A describe the state and the action of t-time of the process ($t \ge 0$). 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 $t \ge 0$ 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 $t \ge 0$. We assume that for each $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$,

$$P^\pi(X_{t+1} = j|H_{t-1}, \Delta_{t-1}, X_t = i, \Delta_t = a) = q_{ij}(a) \tag{2.2}$$

for all $t \ge 0, i, j \in S, a \in A(i)$. 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 $P_\nu^\pi \in \mathcal{P}(\Omega)$ by a usual way.

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

$$\mathcal{B} := \sum_{t=0}^{\infty} \beta^t r(X_t, \Delta_t, X_{t+1}), \tag{2.3}$$

and

$$\mathcal{B}_t := \sum_{k=0}^{t} \beta^k r(X_k, \Delta_k, X_{k+1}) \ \text{ for } \ t \geq 0 \tag{2.4}$$

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

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

$$F_\nu^\pi(z) := P_\nu^\pi(\mathcal{B} \leq z), \tag{2.5}$$

$$\Phi(\nu) := \{F_\nu^\pi(\cdot) | \pi \in \Pi\}. \tag{2.6}$$

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

Now we will derive the optimality equation under arbitrary continuous function g, which constructs a g-optimal policy. By weak-compactness of $\Phi(\nu)$, the following existence theorem holds.

Theorem 2.1. *For any $\nu \in P(S)$ and $g \in \mathcal{C}[0, M_\beta]$, there exists a (ν, g)-optimal policy.*

For simplicity of the notation, let

$$U_t\{g\}(s, i, a, j) := \max_{F \in \Phi(j)} \int_0^{M_\beta} g(s + \beta^t r(i, a, j) + \beta^{t+1} z)\, F(dz) \tag{2.7}$$

for $t \geq 0, g \in \mathcal{C}[0, M_\beta], s \in [0, M_\beta(1-\beta^t)], i, j \in S$ and $a \in A(i)$. And, if $\nu \in P(S)$ is degenerate at $\{j\}$, ν is simply denoted by j and $\Phi(\nu)$ by $\Phi(j)$ similarly.

Now, we can state one of our main results, which gives a necessary condition for (ν, g)-optimality.

Theorem 2.2. *For any $\nu \in P(S)$ and $g \in \mathcal{C}[0, M_\beta]$, let $\pi^* \in \Pi$ be (ν, g)-optimal. Then for each $t \geq 0$, the following optimal equation holds.*

$$E_\nu^{\pi^*}[g(\mathcal{B})] = E_\nu^{\pi^*}[\max_{a \in A(X_t)} \sum_{j \in S} q_{X_t j}(a) U_t\{g\}(\mathcal{B}_{t-1}, X_t, a, j)], \tag{2.8}$$

where $\mathcal{B}_{-1} := 0$.

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

$$A_t^*(s, i) := \arg\max_{a \in A(i)} \sum_{j \in S} q_{ij}(a) U_t\{g\}(s, i, a, j). \tag{2.9}$$

Theorem 2.3. *For any $\nu \in P(S)$ and $g \in \mathcal{C}[0, M_\beta]$, 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.

3 Utility-Optimal Stopping Problem

This section is concerned with a general utility of the optimal stopping problem for denumerable Markov chains. The optimality of the one-step look ahead (OLA) policy is shown under suitable conditions.

As for the utility theory, so many authors analyzed decision processes with it. For such examples of Markov decision processes, see Howard and Matheson[8] and Chung and Sobel[3]. The analysis under a general utility criterion has been done, for example, in Rieder[17] and our previous paper[10], which is expected to enlarge the practical applications of the utility. To our knowledge, Denardo and Rothblum[4] is the only work related with utility treatment of optimal stopping problem. They analyze the problem in a finite Markov decision chain

with the exponential utility and give a linear programming corresponding to an optimal stopping time.

Let us designate a transition law as $Q = (q_{ij}; i, j \in S)$ and the underlying process as $\{X_t\}$ by dropping the action since the decision does not imposed in this section, for a denumerable state space S.

A stopping time is a random variable $\sigma : \Omega = S^\infty \to \{0, 1, 2, \cdots\}$ such $P_\nu(\sigma < \infty) = 1$ and $\{\sigma = t\}$ is measurable with respect to the σ-algebra induced by $\{X_0, X_1, \cdots, X_t\}$ for $t = 0, 1, 2, \cdots$ where ν is an initial distribution on S. Let denote by Σ_ν the set of all stopping times starting with the initial distribution $\nu \in \mathcal{P}(S)$. Let $\mathcal{R}$ be the set of all real numbers. The terminal reward at the state $i \in S$, $r_i = r(i)$, is a function from S to $\mathcal{R}$ and the observation cost per unit time is a constant $c > 0$. The total reward when the system is stopped at time t is given by the random variable

$$\mathcal{B}_t := c_t + r(X_t), \tag{3.1}$$

where $c_t := -ct$.

A utility g is a Borel measurable function from $\mathcal{R}$ to itself. Let denote by $E_\nu[Y]$ the expectation of a random variable Y with respect to P_ν. For any utility g and the initial distribution $\nu \in \mathcal{P}(S)$, our optimal stopping problem is to maximize the expected utility $E_\nu[g(\mathcal{B}_\sigma)]$ over all $\sigma \in \Sigma_\nu$.

For any g and $\nu \in \mathcal{P}(S)$, the stopping time $\sigma^* \in \Sigma_\nu$ is called (ν, g)-*optimal*, if

$$E_\nu[g(\mathcal{B}_{\sigma^*})] \geq E_\nu[g(\mathcal{B}_\sigma)] \tag{3.2}$$

for all $\sigma \in \Sigma_\nu$. The $\sigma^* \in \bigcap_{\nu \in \mathcal{P}(S)} \Sigma_\nu$ is called g-*optimal* if it is (ν, g)-optimal for all $\nu \in \mathcal{P}(S)$.

In the subsequent discussion, it is convenient to rewrite the expected utility $E_\nu[g(\mathcal{B}_\sigma)]$ by using the distribution function of $\mathcal{B}_\sigma$ with respect to P_ν. For this purpose, we define

$$U\{g\}(a, i) := \sup_{F \in \Phi(i)} \int_{-\infty}^{\infty} g(a + x)\, F(dx) \tag{3.3}$$

for $a \in \mathcal{R}$ and $i \in S$, where $\Phi(i)$ denotes $\Phi(\nu)$ for ν such that $\nu(\{i\}) = 1$. This class of $\Phi(\nu)$ is defined similarly as (2.6).

The validity of the OLA stopping times is discussed. In order to characterize the optimal stopping time, we consider the following set and its hitting time.

For each $t = 0, 1, 2, \cdots$, let

$$S_t\{g\} := \{\, i \in S \,|\, g(c_t + r_i) \geq \sum_{j \in S} q_{ij} U\{g\}(c_{t+1}, j) \,\} \tag{3.4}$$

and

$$\sigma^* := \{ \text{ the first time } t \text{ such that } X_t \in S_t\{g\}\}\,. \tag{3.5}$$

The next assumption is fundamental for a general theory in the optimal stopping.

Assumption 3.1. *For any $\nu \in \mathcal{P}(S)$,*

$$E_\nu[\sup_{\{t \geq 0\}} g(\mathcal{B}_t)^+\,] < \infty. \tag{3.6}$$

By applying the result in Chow, Robbins and Siegmund[2] to the sequence of random variables $\{g(\mathcal{B}_t)\}_{t=0,1,2,\ldots}$, we can show the next theorem.

Theorem 3.2. (refer Theorem 4.5 in [2])

(i) *Suppose* Assumption 3.1 *and* $P_\nu(\,\sigma^* < \infty\,) = 1$ *for any* $\nu \in \mathcal{P}(S)$. *Then,* $\sigma^* \in \Sigma_\nu$ *and* σ^* *is* (ν, g)*-optimal.*

(ii) *Suppose that* Assumption 4.1 *and that* $\lim_{t\to\infty} g(\mathcal{B}_t) = -\infty$ P_ν*-a.s. holds for any* $\nu \in P(S)$. *Then,* $\sigma^* \in \bigcap_{\nu \in \mathcal{P}(S)} \Sigma_\nu$ *and* σ^* *is g-optimal.*

Now, using an idea of the OLA stopping time for optimal stopping problems with additive utility functions (for example, see [18]), we derive some results on the general utility case. For each t, let

$$S_t^*\{g\} := \{\, i \in S \,|\, g(c_t + r_i) \geq \sum_{j \in S} q_{ij}\, g(c_{t+1} + r_j)\}. \tag{3.7}$$

Notice, from (3.3) and (3.4), that $S_t\{g\} \subset S_t^*\{g\}$ for all t. The OLA stopping time is a stopping time whose value is determined by the first hitting time t such that $X_t \in S_t^*\{g\}$. Here, we introduce an assumption to get useful results on the validity of the OLA stopping time.

Assumption 3.3. *For each* $t = 0, 1, 2, \cdots$, $Q = (q_{ij})$ *and* $S_t^*\{g\}$ *satisfy that*

$$q_{ij} = 0 \text{ if } i \in S_t^*\{g\} \text{ and } j \notin S_{t+1}^*\{g\}. \tag{3.8}$$

If $S_0^*\{g\} = S_t^*\{g\}$ for all t, Assumption 3.3 assures the *closedness* of $S_0^*\{g\}$. Notice that if $S_t^*\{g\} \neq \emptyset$ for some t, then $S_{t+n}^*\{g\} \neq \emptyset$ for $n = 1, 2, \cdots$.

Let denote E_ν and Σ_ν for $\nu \in \mathcal{P}(S)$ such that $\nu(\{i\}) = 1$ by E_i and Σ_i, respectively. In the following lemma, the relation (3.10) is sometimes called the monotone property.

Assumption 3.4. *Suppose that*

$$E_i[\sup_{\{k\geq 0\}} g(c_t + \mathcal{B}_k)^+] < \infty \tag{3.9}$$

for any $i \in S_t^*\{g\}$ *and* $t = 0, 1, 2, \cdots$.

Lemma 3.5. *Assumption 3.3 and 3.4 imply that*

$$U\{g\}(c_t, j) = g(c_t + r_j) \quad \text{for any } j \in S_t^*\{g\}\, t = 0, 1, 2, \cdots. \tag{3.10}$$

The next theorem gives a sufficient condition for the OLA stopping time to be optimal under the general utility.

Theorem 3.6. *If Assumption 3.3 and 3.4 hold, then it holds that* $S_t^*\{g\} = S_t\{g\}$ *for* $t = 0, 1, 2, \cdots$.

In case of a linear utility function $g(x) = x$, it is reduced to

$$S^* := S_t^*\{x\} = \{\, i \in S \,|\, c + r_i \geq \sum_{j\in S} q_{ij} r_j \}, \tag{3.11}$$

which is independent of t and so we denote it as S^*. The next theorem shows a property of the OLA stopping times characterized by the non-decreasing utility.

Theorem 3.7. *Let* g *be a non-decreasing function.*

(i) *If it is concave, then* $S_t^*\{g\} \supset S^*$ *for each* t.
(ii) *If it is convex, then* $S_t^*\{g\} \subset S^*$ *for each* t.

We note that the concave function is risk-averse and the convex one is risk-seeking. The OLA stopping time of a risk-averse decision maker has a tendency to stop earlier than that of a risk seeking one.

4 Stopped decision processes with general utility

In this section we will discuss the problem, so called as stopped decision process by Furukawa[6], Furukawa and Iwamoto[7], under the frame work of general utility.

Following the notation of Section 2, the standard Markov decision process

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

are assumed to be given as (2.1). Let define a stopping time $\sigma : \Omega \to \{0, 1, 2, \cdots\}$ by the following condition:

(i) *For each* n, $\{\sigma = n\} \in \mathcal{F}_n$,
(ii) $P_\nu^\pi(\sigma < \infty) = 1$,
(iii) $E_\nu^\pi[g(\mathcal{B}_\sigma)^-] < \infty$,

where $\nu \in \mathcal{P}(S), \pi \in \Pi$ and $\{\mathcal{F}_n\}$ is a given σ-algebra. For each $\nu \in P(S), \pi \in \Pi$, we call this a stopping time σ with respect to (ν, π) with a fixed g. Denote the class of a stopping time σ and that of a pair (π, σ) by

$$\Sigma_{(\nu,\pi)} := \{ \text{ a stopping time } \sigma \text{ w.r.t. } (\nu, \pi)\} \tag{4.1}$$

and

$$\mathcal{A}_\nu := \{(\pi, \sigma) | \sigma \in \Sigma_{(\nu,\pi)}, \pi \in \Pi\} \tag{4.2}$$

respectively.

Our problem is to find a pair of a policy π and a stopping time σ which maximize the expected utility;

$$E_\nu^\pi[g(\mathcal{B}_\sigma)] = E_\nu^\pi\left[g(\Sigma_{t=0}^\sigma \beta^{t-1} r(X_{t-1}, \Delta_{t-1}, X_t))\right]$$

over $(\pi, \sigma) \in \mathcal{A}_\nu$ for each $\nu \in P(S)$ where $r(X_{-1}, \Delta_{-1}, X_0) = 0$.

Definition 4.1. *The pair of* $(\pi^*, \sigma^*) \in \mathcal{A}_\nu$ *is* (ν, g)*-optimal or simply* ν*-optimal if*

$$E_\nu^{\pi^*}[g(\mathcal{B}_{\sigma^*})] \geq E_\nu^\pi[g(\mathcal{B}_\sigma)] \tag{4.3}$$

for all $(\pi, \sigma) \in \mathcal{A}_\nu$.

To consider an optimality equation, define

$$U_t\{g\}(s, i) := \max_{F\in\Phi(i)} \int_{-\infty}^{\infty} g(s + \beta^t z)\, F(dz)$$

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

where $\Phi(i) := \Phi(\{i\}), \Phi(\nu) := \{F^{\nu}_{\pi,\sigma} \mid (\pi,\sigma) \in \mathcal{A}\}$, $F^{\nu}_{\pi,\sigma}(z) := P^{\nu}_{\pi}(\mathcal{B}_{\sigma} \leq z)$ as similar to (2.7). However we note that this class of distributions is not compact since the domain is not bounded.

Assumption 4.2.

(i) *For each* $i, j \in S$, *we assume that* $q_{ij}(\cdot)$ *and* $r(i,\cdot,j)$ *are continuous in* $a \in A(i)$.
(ii) *The general utility function* $g(x), x \in S$ *is differentiable and* $\mid g'(x) \mid$ *is uniformly bounded.*
(iii) *The initial distribution* $\nu \in P(S)$ *has a finite support.*
(iv) $E^{\pi}_{\nu}[\sup_{t\geq 0} g(\mathcal{B})^{+}] < \infty$.

We note that this assumption is satisfied when it is a neutral or a risk averse case.

Lemma 4.3. *Under* Assumption 4.2, *the following* (i) *and* (ii) *are hold.*

(i) $U_t\{g\}(s,i)$ *is continuous in* $s \in R$ *for each* $i \in S$.
(ii) $\Sigma_{j\in S} q_{ij}(a) U_t\{g\}(s,i,a,j)$ *is continuous in* $a \in A(i)$ *for each* $i \in S$ *and* t.

In order to characterize the optimal pair in the recursive form, let define a sequence of sets of the pairs by a truncation of the time horizon.

Definition 4.4. For $n \geq 1$,

$$\mathcal{A}^{n}_{\nu} := \{\, (\pi, n \vee \sigma) \,|\, (\pi,\sigma) \in \mathcal{A}_{\nu} \}. \tag{4.4}$$

Note that $\mathcal{A}^{1}_{\nu} = \mathcal{A}_{\nu}$ and $\mathcal{A}^{1}_{\nu} \supset \mathcal{A}^{2}_{\nu} \supset \mathcal{A}^{3}_{\nu} \supset \cdots$ holds clearly. By using a shift operator θ_n for a history $H_t = (X_0, \Delta_0, X_1, \cdots, X_t)$, let $\theta_n H_t := (X_n, \Delta_n, X_{n+1}, \cdots, X_t)$ for $n \leq t$ and $\theta_t H_t := X_t$. A truncated policy π^n at n-th is

$$\pi^n := (\pi_0, \pi_1, \cdots, \pi_{n-1}) \tag{4.5}$$

for $\pi = (\pi_0, \pi_1, \cdots)$ and $(\pi,\sigma) \in \mathcal{A}^{n}_{\nu}$. A conditional pair is the pair of a policy $\pi[H_n] := (\pi[H_n]_1, \pi[H_n]_2, \cdots)$ and a conditional stopping time $\sigma[H_n](\omega) := \sigma(\omega') - (n-1)$ with a given history H_n where $\omega = (X'_1, \Delta'_1, \cdots)$ and $\omega' = (H_n, X'_1, \Delta'_1, \cdots) \in \Omega$. Let us denote that $\mathcal{B}'_n := \sum_{t=0}^{n} \beta^t r(X'_t, \Delta'_t, X'_{t+1})$ for $n \geq 0$.

Lemma 4.5. *If* $(\pi,\sigma) \in \mathcal{A}^{n}_{\nu}$, *then*

$$(\pi[H_n], \sigma[H_n]) \in \mathcal{A}_{X_n} \quad P^{\pi}_{\nu}\text{-a.s.} \tag{4.6}$$

where $\mathcal{A}_{X_n}$ *means the initial distribution* ν *of* $\mathcal{A}_{\nu}$ *is degenerated at* X_n.

Proof. Assume contrarily that $P_\nu^\pi((\pi[H_n], \sigma[H_n]) \notin \mathcal{A}_{X_n}) > 0$. Since $(\pi, \sigma) \in \mathcal{A}_\nu^n$,

$$\begin{aligned} 1 &= P_\nu^\pi(\sigma < \infty) \\ &= P_\nu^\pi(n \le \sigma < \infty) \\ &= E_\nu^\pi[P_\nu^\pi(n \le \sigma < \infty) \mid \mathcal{F}_n] \\ &= E_\nu^\pi[P_{X_n}^{\pi[H_n]}(\sigma[H_n] < \infty)] \\ &= E_\nu^\pi[P_{X_n}^{\pi[H_n]}(\sigma[H_n] < \infty)\mathbf{1}_D] + P_\nu^\pi(D^c). \end{aligned}$$

The last term is divided to the two cases such that $D := \{(\pi[H_n], \sigma[H_n]) \notin \mathcal{A}_{X_n}\}$ and its negation D^c using the indicator function $\mathbf{1}$. The assumption that $0 \le P_\nu^{\pi[H_n]}(\sigma[H_n] < \infty) \le 1$ and $P_\nu^\pi(D) > 0$ imply $P_\nu^{\pi[H_n]}(\sigma[H_n] < \infty) = 1$, $P_\nu^{\pi[H_n]}$−a.s. on D. On the other hand, using by Assumption 4.2(ii), there exist a constant M_n and K such that $|\mathcal{B}_n'| \le M_n$ and $g(\mathcal{B}_n + z) \le g(z) + KM_n$ for all z. Using this evaluation and $E_\nu^\pi[g(\mathcal{B}_\sigma)^-] < \infty$, we obtain that $E_\nu^\pi[E_{X_n}^{\pi[H_n]}(g(\mathcal{B}'_{\sigma[H_n]})^-] < \infty$. So $P_\nu^\pi((\pi[H_n], \sigma[H_n]) \notin \mathcal{A}_{X_n}) = 0$ follows. This completes the proof.

For the rest of assertion its proof is omitted for sake of shrinking the pages.

Lemma 4.6. *For an arbitrary initial state $i \in S$ and $\pi \in \Pi$, we select a pair $(\pi^i, \sigma^i) \in \mathcal{A}_i$. Define (π', σ') by $\pi' := (\pi^n, \pi'_{n+1}, \pi'_{n+2}, \cdots)$ with $\pi'_{n+k}(H_{n+k}) = \pi^{X_n}(\theta_n H_{n+k})$ $(k \ge 1)$ and $\sigma'(\omega) := \sigma^{X_n}(\theta_n\omega) + n$ for any ω in the sample space and any fixed $n \ge 1$. Then*

$$(\pi', \sigma') \in \mathcal{A}_\nu^n \tag{4.7}$$

provided Assumption 4.2 *holds.*

Definition 4.7. *Define a conditional maximum reward by*

$$\gamma_n^\nu := esssup_{(\pi,\sigma)\in\mathcal{A}_\nu^n} E_\nu^\pi[g(\mathcal{B}_\sigma) \mid \mathcal{F}_n] \quad (n \ge 1) \tag{4.8}$$

Henceforth we write esssup by sup for a short.

Lemma 4.8. *For each $n \ge 1$,*

(i) $\gamma_n = U\{g\}(\mathcal{B}_n, X_n)$
(ii) $\gamma_n = \max\{g(\mathcal{B}_n), \sup_{\pi\in\Pi} E_\nu^\pi(\gamma_{n+1}|\mathcal{F}_n)\}$
(iii) $\sup_{\pi\in\Pi} E_\nu^\pi(\gamma_{n+1}|\mathcal{F}_n)\}$
$= \max_{a\in A(X_n)} \Sigma_{j\in S} q_{X_n,j}(a) U_n\{g\}(\mathcal{B}_n, X_n, a, j)$

where $\gamma_n = \gamma_n^\nu$ for a sake of simplicity.

Let $\sigma^* =$ the first time t such that $g(\mathcal{B}_t) = \sup_{\pi\in\Pi} E_\nu^\pi(\gamma_{t+1}|\mathcal{F}_t)\}$ and define, for $s \in R$, $i \in S$,

$$A_t^*(s,i) := \text{argmax}_{a\in A(i)} \Sigma_{j\in S} q_{ij} U_t\{g\}(s,i,a,j). \tag{4.9}$$

Theorem 4.9. *If a policy $\pi^* = (\pi_1^*, \pi_2^*, \cdots)$ satisfies*

$$\pi_n^*(A_n^*(\mathcal{B}_n, X_n) \mid H_n) = 1 \tag{4.10}$$

for $n \geq 1$, the next iterative relation

$$\gamma_n = \max\{g(\mathcal{B}_n), E_\nu^{\pi^*}[\gamma_{n+1} \mid \mathcal{F}_n]\} \quad (n \geq 1) \tag{4.11}$$

is obtained.

Lemma 4.10. *For π^* satisfying (4.10),*

$$\{\gamma_{n\wedge\sigma^*}, \mathcal{F}_n; n = 1, 2, \cdots\} \tag{4.12}$$

is a martingale with respect to $P_\nu^{\pi^*}$.

Theorem 4.11. *Assume the previous Assumption 4.2 and (4.10). Then;*

(i) *If $P_\nu^{\pi^*}(\sigma^* < \infty) = 1$, then (π^*, σ^*) is a ν-optimal pair.*
(ii) *The above condition* (i) *is satisfied when*

$$g(\mathcal{B}_n) \to -\infty \,(\text{as } n \to \infty) \quad P_\nu^{\pi^*}\text{-a.s.}$$

Example 4.12. (Markov deteriorating system and exponential utility)

A simple example for Markov deteriorating system and exponential utility is illustrated. This system is formulated by the following conditions:

(i) *The feasible action space is independent of $i \in S$, that is, $A(i) = A$ for a compact set A.*
(ii) *The utility function is an exponential type; $g(x) := 1 - e^{-\lambda x} (\lambda > 0)$.*
(iii) *For the transition probability, $P_{ij}(a) = 0, i > j$ for all $a \in A$ and $i, j \in S$.*
(iv) *For the reward function, $r(i', a, j) \leq r(i, a, j), i \leq i' \leq j$ for all $i, i', j \in S$.*

Under these assumptions, we immediately calculate the optimal stopping time by using the monotone case of Chow, Robbins and Siegmund[2] and the OLA policy by Ross[18].

Let

$$k_t := \min\{i \geq 0 | \min_{a\in A} \sum_j q_{ij}(a) \exp\{-\lambda\beta^t r(i,a,j)\} \geq 1\} \quad (t \geq 0).$$

Then $k_0 \geq k_1 \geq k_2 \geq \cdots$. The optimal stopping time σ^* defined by the first hitting time t such that $X_t \geq k_t$, which is independent of $\mathcal{B}_t$, is finite $P_i^{\pi^*}$-a.s. for each $i \in S$ and the initial distribution ν of a finite support. The optimal policy could not described explicitly in this situation but it is possible numerically because the optimal stopping region have been determined.

Acknowledgments. The authors are grateful to two anonymous referees for their correcting mistakes and improving earlier version of the article.

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 and Optimization*, **25**(1987), 49-62.

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

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

[6] N.Furukawa, Functional Equations and Markov Potential Theory in Stopped Decision Processes, *Mem. Fac. Sci, Kyushu Univ.*, **29**(1972), 329–347.

[7] N. Furukawa and S. Iwamoto, Stopped Decision Processes on Complete Separable Metric Spaces, *J. Math. Anal. Appl.*, **31**(1970), 615–658.

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

[9] S.C. Jaquette, Markov decision processes with a new optimality criterion: Discrete time, *Ann. Stat.*, **1**(1973), 496–505.

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

[11] Y. Kadota, M. Kurano and M. Yasuda, Utility-Optimal Stopping in a denumerable Markov chain, *Bulletin of Information and Cybernetics,* **28**(1996), 15 – 21.

[12] M. Kurano, Markov decision processes with a minimum-variance criterion, *J. Math. Anal. Appl.*, **123** (1987), 572–583.

[13] P. Mandl, On the variance of controlled Markov chains, *Kybernetika*, **7** (1971), 1–12.

[14] L.T. Nielsen, The expected utility of portfolios of assets, *J. Math. Economics*, **22** (1993), 439–461.

[15] E.L. Porteus, On the optimality of structured policies in countable stage decision processes, *Manag. Sci.*, **22** (1975), 148–157.

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

[17] U. Rieder, *Non-cooperative dynamic games with general utility functions*, T. E. S. Raghavan et al.(eds). Stochastic Games and Related Topics, Klumer Academic publishers. 161–174, 1991.

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

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

[20] N. L. Stokey and R. E. Lucas, Jr, Recursive methods in economic dynamics, Harvard University Press, 1989.

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