American Journal of Computational Mathematics, **2011, 1, 183-188**
doi:10.4236/ajcm.2011.13021 Published Online September 2011 (http://www.SciRP.org/journal/ajcm)

Conditional Value-at-Risk for Random Immediate Reward Variables in Markov Decision Processes

Masayuki Kageyama, Takayuki Fujii, Koji Kanefuji, Hiroe Tsubaki
The Institute of Statistical Mathematics, *Tokyo*, *Japan*
E-mail: *kageyama@ism.ac.jp*
Received May 17, 2011; *revised August* 10, 2011; *accepted August* 22, 2011

Abstract

We consider risk minimization problems for Markov decision processes. From a standpoint of making the risk of random reward variable at each time as small as possible, a risk measure is introduced using conditional value-at-risk for random immediate reward variables in Markov decision processes, under whose risk measure criteria the risk-optimal policies are characterized by the optimality equations for the discounted or average case. As an application, the inventory models are considered.

Keywords: Markov Decision Processes, Conditional Value-at-Risk, Risk Optimal Policy, Inventory Model

1. Introduction

As a measure of risk for income or loss random variables, the variance has been commonly considered since Markowitz work [1]. The variance has the shortcoming that it does not approximately account for the phenomenon of "fat tail" in distribution functions. In recent years, many risk measures have been generated and analyzed by an economically motivated optimization problem, for example, value at risk $(V@R)$, conditional value-at-risk $(CV@R)$ [2,3], coherent risk of measure [4-6], convex risk of measure [7,8] and its applications [9,10].

On the other hand, a lot of research considering the risk have been progressed by many authors [11-15] in the framework of Markov decision processes (MDPs, for short). In [11,16], the risk control for the random total reward in MDPs is discussed. In the sequential decision making under uncertain circumstance, it may be better to minimize the total risk through the infinite horizon controlling the risk at each time. For example, in multiperiod inventory and production problem, we often want to order optimally by the ordering policy such that while it minimizes the total risk through all the periods it also makes the risk at each time as small as possible.

In this paper, with above motivation in mind we introduce a new risk measure for each policy using conditional value-at-risk for random immediate reward variables, under whose risk measure criteria the optimization will be done, respectively, in the discounted and average case. As an application, the inventory model is considered. In the reminder of this section, we shall establish notations that will be used throughout the paper and define the problem with a new risk measure.

A Borel set is a Borel subset of a complete separable metric space. For a Borel set X, B_X denotes the $\sigma-$ algebra of Borel subset of X. For Borel sets X and Y, $\mathrm{P}(X)$ and $\mathrm{P}(X|Y)$ be the sets of all probability measures on X and all conditional probability measures on X given Y respectively. The product of X and Y is denoted by XY. Let $\mathbb{R}$ be the set of real numbers. Let I be a random income (or reward) variable on some probability space (Ω, B, P), and $F_I(x)$ the distribution function of I, *i.e.*, $F_I(x) = P(I \le x)(x \in \Re)$. We define the inverse function $F_I^{-1}(p)(0 \le p \le 1)$ by

$$F_I^{-1}(p) = \inf\{x \in \Re | F_I(x) \ge p\}.$$

Then, the Conditional Value-at- Risk for a level $\gamma \in (0,1)$ of I, $CV@R_\gamma(I)$, is defined (cf. [2,3]) by

$$CV@R_\gamma(I) = \frac{1}{1-\gamma}\int_\gamma^1 F_{-I}^{-1}(p)\mathrm{d}p. \qquad (1)$$

We note that $CV@R_\gamma(I)$ is specified depending only on the law of the random variable I. For any Borel set X, the set of all bounded and Borel measurable functions on X will be denoted by $\mathrm{B}(X)$.

A Markov decision process is a controlled dynamic system defined by a six-tuple $\{S, A, \{A(x) | x \in A\}, \mathrm{Q}, \tilde{r}, \nu\}$, where Borel sets S and A are state and action spaces, respectively, $A(x)$ is non-empty Borel subset of A which denotes the set of feasible actions when the system

is in state $x \in S$, $Q \in \mathrm{P}(S|SA)$ is the law of motion, $\tilde{r} \in \mathrm{B}(SAS)$ is an immediate reward function and $\nu \in \mathrm{P}(S)$ is an initial state distribution.

Throughout this paper, we suppose that the set

$$K := \{(x,a) \in SA | a \in A(x) \text{ for } x \in S\}$$

is in B_{SA}. The sample space is the product space $\Omega = (SA)^{\infty}$ such that the projections X_t, Δ_t on the *t*-th factors S, A describe the state and the action at the *t*-th time of the process $(t \geq 0)$.

Let Π denotes the set of all policies, *i.e.*, for $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$ let $\pi_t \in \mathrm{P}\left(A \middle| S(AS)^t\right)$ with

$$\pi_t(A(x_t) | x_0, a_0, \cdots, a_{t-1}, x_t) = 1$$

for all $(x_0, a_0, \cdots, a_{t-1}, x_t) \in S(AS)^t (t \geq 0)$. If there is a Borel measurable function $f : S \to A$ with $f(x) \in A(x)$ for all $x \in S$ such that $\pi_t(\{f(x_t)\} | x_0, a_0, \cdots, a_{t-1}, x_t) = 1$ for all $(x_0, a_0, \cdots, a_{t-1}, x_t) \in S(AS)^t (t \geq 0)$, a policy $\pi = (\pi_0, \pi_1, \cdots)$ is called stationary. Such a policy will be denoted by *f*. Let $H_t = (X_0, \Delta_0, \cdots, X_t, \Delta_t)$. For any $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$, we assume that

$$Pr(\Delta_t \in \mathrm{D}_1 | H_{t-1}, X_t) = \pi_t(\mathrm{D}_1 | H_{t-1}, X_t) \tag{2}$$

and

$$\begin{aligned} &Pr(X_{t+1} \in \mathrm{D}_2 | H_{t-1}, X_t = x, \Delta_t = a) \\ &= Q(\mathrm{D}_2 | x, a) \end{aligned} \tag{3}$$

for $\mathrm{D}_1 \in \mathrm{B}_A, \mathrm{D}_2 \in \mathrm{B}_S, x \in S, a \in A(x)$ and $t \geq 0$. Then, for any $\pi \in \Pi$ and initial state distribution $\nu \in \mathrm{P}(S)$, the probability measure $P_\pi^\nu(\cdot)$ is given on Ω in an obvious way. If not specified otherwise, P_π^ν is denoted by P_π suppressing ν in P_π^ν.

We want to minimize the total reward risk making the risk at each time as small as possible. So, using *CV@R* for the random reward variable $\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t)$ at time *t*, a risk measure $\rho(\tilde{r}|\pi)$ for the random reward stream $\{\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t) : t = 1, 2, \cdots\}$ will be defined in the discounted or average case as follows. With some abuse of notation, we denote by $\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t) | H_{t-1}$ the conditional distribution of $\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t)$ given $H_{t-1} (t \geq 1)$. Also, E_π is the expectation operator w.r.t. P_π.

a) The discounted case $(0 < \beta < 1)$.

$$\begin{aligned} \rho_{DS}(\tilde{r}|\pi) := \frac{1}{1-\beta} \sum_{t=1}^{\infty} \beta^t E_\pi \\ \left[CV@R_\gamma \left(\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t) | H_{t-1} \right) \right]. \end{aligned} \tag{4}$$

b) The average case.

$$\rho_{AV}(\tilde{r}|\pi) := \limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E_\pi \left[CV@R_\gamma \left(\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t) | H_{t-1} \right) \right]. \tag{5}$$

For the family of random reward streams

$$\left\{ \left(\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t) : t = 1, 2, \cdots \right) : \tilde{r} \in \mathrm{B}(SAS) \right\},$$

$\rho_{DS}(\tilde{r}|\pi)$ and $\rho_{AV}(\tilde{r}|\pi)$ have same properties as those of coherent risk measures (cf. [4]), which is shown in the following proposition.

Proposition 1.1. *For any* $\pi \in \Pi$, ρ_{DS} *and* ρ_{AV} *have the following* 1) - 4)*:*

1) (Monotonicity) If $\tilde{r}_1 \leq \tilde{r}_2$ with $\tilde{r}_1, \tilde{r}_2 \in \mathrm{B}(SAS)$, $\rho(\tilde{r}_1) \geq \rho(\tilde{r}_2)$.

2) (Translation invariance) For $\tilde{r} \in \mathrm{B}(SAS)$ and $c \in \Re = (-\infty, \infty)$, $\rho(\tilde{r} + c) = \rho(\tilde{r}) - c$.

3) (Homogeneity) For $\tilde{r} \in \mathrm{B}(SAS)$ and $\lambda > 0$, $\rho(\lambda \tilde{r}) = \lambda \rho(\tilde{r})$.

4) (Convexity) For $\tilde{r}_1, \tilde{r}_2 \in \mathrm{B}(SAS)$ and $0 \leq \lambda \leq 1$, $\rho(\lambda \tilde{r}_1 + (1-\lambda)\tilde{r}_2) \leq \lambda \rho(\tilde{r}_1) + (1-\lambda)\rho(\tilde{r}_2)$.

Proof. Notice that

$$\rho(\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t) | H_{t-1}) = CV@R_\gamma(\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t) | H_{t-1})$$

satisfies the properties 1)-4) for $\tilde{r} \in \mathrm{B}(SAS)$ For 4) with $\rho(\cdot) = \rho_{AV}(\cdot|\pi)$, suppose that $\tilde{r}_1, \tilde{r}_2 \in \mathrm{B}(SAS)$ and $0 \leq \lambda \leq 1$. Then, we have that

$$\begin{aligned} &\rho_{AV}\left((\lambda \tilde{r}_1 + (1-\lambda)\tilde{r}_2 | H_{t-1}) | \pi \right) \\ &= \limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E_\pi \left[CV@R_\gamma \left((\lambda \tilde{r}_1 + (1-\lambda)\tilde{r}_2) | H_{t-1} \right) \right] \\ &= \limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E_\pi \left[CV@R_\gamma \left(\lambda(\tilde{r}_1 | H_{t-1}) \right. \right. \\ &\qquad \left. \left. + (1-\lambda)(\tilde{r}_2 | H_{t-1}) \right) \right] \\ &\leq \limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E_\pi \left[\lambda CV@R_\gamma (\tilde{r}_1 | H_{t-1}) \right. \\ &\qquad \left. + (1-\lambda) CV@R_\gamma (\tilde{r}_2 | H_{t-1}) \right], \end{aligned}$$

from convexity of *CV@R*,

$$\begin{aligned} &\leq \lambda \limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E_\pi \left[CV@R_\gamma (\tilde{r}_1 | H_{t-1}) \right] \\ &+ (1-\lambda) \limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E_\pi \left[CV@R_\gamma (\tilde{r}_2 | H_{t-1}) \right] \\ &= \lambda \rho_{AV}(\tilde{r}_1 | \pi) + (1-\lambda) \rho_{AV}(\tilde{r}_2 | \pi), \end{aligned}$$

which implies (iv) with $\rho = \rho_{AV}$. Other assertions in Proposition 1.1 are easily proved. This completes the proof. □

For $\tilde{r} \in \mathrm{B}(SAS)$, and $(x,a) \in K$, the conditional distribution function $D_{-\tilde{r}}(\cdot | x, a)$ is defined by

$$D_{-\tilde{r}}(y | x, a) := Q\left(\bigcup (y | x, a, \tilde{r}) | x, a \right), \tag{6}$$

where $\bigcup (y | x, a, \tilde{r}) := \{ z \in S | -\tilde{r}(x, a, z) \leq y \}$.

Lemma 1.2. *For any* $\pi \in \Pi$ *it holds that*

$$\begin{aligned} & E_\pi\left[CV@R_\gamma\left(\tilde{r}\left(X_{t-1},\Delta_{t-1},X_t\right)\middle|H_{t-1}\right)\right] \\ &= E_\pi\left[CV@R_\gamma\left(\tilde{r}\left(X_{t-1},\Delta_{t-1},X_t\right)\middle|X_{t-1},\Delta_{t-1}\right)\right] \\ &= E_\pi\Big[D_{-\tilde{r}}^{-1}\left(\gamma\middle|X_{t-1},\Delta_{t-1}\right) \\ &\quad + \frac{1}{1-\gamma}\int\left[-\tilde{r}\left(X_{t-1},\Delta_{t-1},y\right)-D_{-\tilde{r}}^{-1}\left(\gamma\middle|X_{t-1},\Delta_{t-1}\right)\right]^+ \\ &\quad \mathrm{Q}\left(\mathrm{d}y\middle|X_{t-1},\Delta_{t-1}\right)\Big], \end{aligned} \tag{7}$$

where $[x]^+ := \max\{x,0\}$.

Proof. From the Markov property (3), it follows that

$$\begin{aligned} & P_\pi\left(\tilde{r}\left(X_{t-1},\Delta_{t-1},X_t\right)\le y\middle|H_{t-1}\right) \\ &= P_\pi\left(\tilde{r}\left(X_{t-1},\Delta_{t-1},X_t\right)\middle|X_{t-1},\Delta_{t-1}\right). \end{aligned}$$

Thus,

$$\begin{aligned} & E_\pi\left[CV@R_\gamma\left(\tilde{r}\left(X_{t-1},\Delta_{t-1},X_t\right)\middle|H_{t-1}\right)\right] \\ &= E_\pi\left[CV@R_\gamma\left(\tilde{r}\left(X_{t-1},\Delta_{t-1},X_t\right)\middle|X_{t-1},\Delta_{t-1}\right)\right]. \end{aligned}$$

By the representation formula of $CV@R_\gamma$ (cf. [2,3]), the second equality of (7) holds, which completes the proof. □

The value function of the discounted and average cases are defined respectively by

$$\begin{aligned} \rho_{DS}\left(\tilde{r}\right) &:= \inf_{\pi\in\Pi}\rho_{DS}\left(\tilde{r}\middle|\pi\right) \quad \text{and} \\ \rho_{AV}\left(\tilde{r}\right) &:= \inf_{\pi\in\Pi}\rho_{AV}\left(\tilde{r}\middle|\pi\right) \end{aligned} \tag{8}$$

A policy $\pi^* \in \Pi$ is called discounted and average risk-optimal, respectively, if $\rho_{DS}\left(\tilde{r}\right) = \rho_{DS}\left(\tilde{r}\middle|\pi^*\right)$ and $\rho_{AV}\left(\tilde{r}\right) = \rho_{AV}\left(\tilde{r}\middle|\pi^*\right)$.

2. Risk-Optimization

In this section, using *CV@R* for a random reward variable (1), we define a new immediate reward function by which the theory of MDPs will be easily applicable. Moreover, sufficient conditions are given for the existence of discounted or average risk optimal policies.

2.1. Another Representation of Risk Measures

In this subsection, another representation for ρ_{DS} and ρ_{AV} are given.

For any $\tilde{r} \in \mathrm{B}(SAS)$, the corresponding immediate reward function $r \in \mathrm{B}(SA)$ will be defined by

$$\begin{aligned} r(x,a) = D_{-\tilde{r}}^{-1}\left(\gamma\middle|x,a\right) + \frac{1}{1-\gamma}\int\Big[-\tilde{r}\left(x,a,y\right) \\ -D_{-\tilde{r}}^{-1}\left(\gamma\middle|x,a\right)\Big]^+ \mathrm{Q}\left(\mathrm{d}y\middle|x,a\right) \end{aligned} \tag{9}$$

for each $x \in S$ and $a \in A$. Then, we have the following, which shows that the original problem with $\tilde{r}$ is equivalent to the new problem with r.

Theorem 2.1. *It holds that*, *for any* $\pi \in \Pi$,

1) $\rho_{DS}\left(\tilde{r}\middle|\pi\right) = \dfrac{1}{1-\beta}\sum_{t=0}^{\infty}\beta^t E^\pi\left[r\left(X_t,\Delta_t\right)\right]$

2) $\rho_{AV}\left(\tilde{r}\middle|\pi\right) = \limsup_{T\to\infty}\dfrac{1}{T}\sum_{t=0}^{T-1}E^\pi\left[r\left(X_t,\Delta_t\right)\right]$.

Proof. By Lemma 1.2, it holds that for any $\pi \in \Pi$

$$\begin{aligned} & E_\pi\left[CV@R_\gamma\left(\tilde{r}\left(X_{t-1},\Delta_{t-1},X_t\middle|H_t\right)\right)\right] \\ &= E_\pi\left[r\left(X_{t-1},\Delta_{t-1}\right)\right], (t\ge 1). \end{aligned}$$

So observing the definitions of ρ_{DS} and ρ_{AV} in (4) - (5), 1) and 2) follow, as required. □

2.2. The Discounted Case

Here, we drive the optimality equation for the discounted case, which characterizes a discount risk optimal policy. To this end, we need the following Assumption A.

Assumption A. The following 1) - 4) holds:

1) A is compact and $A(x)$ is closed for each $x \in A$.

2) $\tilde{r}\left(x,a,y\right) \in \mathrm{B}\left(SAS\right)$ is continuous in $(x,a,y) \in SAS$.

3) $\mathrm{Q}\left(\partial\bigcup\left(y\middle|x,a,\tilde{r}\right)\middle|x,a\right) = 0$ for each $(x,a) \in K$ and $y \in \Re$, where

$$\partial\bigcup\left(y\middle|x,a,\tilde{r}\right) = \left\{z \in S\middle|-\tilde{r}\left(x,a,z\right) = y\right\}.$$

4) $\mathrm{Q}\left(\cdot\middle|x,a\right)$ is strongly continuous in $(x,a) \in K$, *i.e.*, for any $v \in \mathrm{B}(S)$, $\int v(y)\mathrm{Q}\left(\mathrm{d}y\middle|x,a\right)$ is continuous in $(x,a) \in K$

Lemma 2.2 *Suppose that Assumption A holds. Then*, $D_{-\tilde{r}}^{-1}\left(\gamma\middle|x,a\right)$ *defined in* (6) *is continuous in* $(x,a) \in K$ *for* $\gamma \in (0,1)$.

Proof. Let $\bigcup^0\left(y\middle|x,a,\tilde{r}\right) := \left\{z \in S\middle|-\tilde{r}\left(x,a,z\right) < y\right\}$. First, we prove that $D_{-\tilde{r}}^{-1}\left(\gamma\middle|x,a\right)$ is lower semi-continuous in $(x,a) \in K$ To the end, it suffices to show that $\underline{D} := \left\{(x,a) \in SA\middle|D_{-\tilde{r}}^{-1}\left(\gamma\middle|x,a\right) \le d\right\}$ is closed for any $d \in \Re$. We observe that $(x,a) \in \underline{D}$ iff for any $\varepsilon > 0$ there exists $y \in \Re$ such that $\mathrm{Q}\left(\bigcup\left(y\middle|x,a,\tilde{r}\right)\middle|x,a\right) \ge \gamma$ and $y \le d + \varepsilon$. Now, let a sequence $\left\{(x_n,a_n) : n = 1,2,\cdots\right\}$ be such that $(x_n,a_n) \in \underline{D}$ and $x_n \to x, a_n \to a\left(n \to \infty\right)$ with $(x,a) \in K$. Then, for any $\varepsilon > 0$, there exists a sequence $\{y_n\}$ with

$$\mathrm{Q}\left(\bigcup\left(y_n\middle|x_n,a_n,\tilde{r}\right)\middle|x_n,a_n\right) \ge \gamma \text{ and } y_n \le d + \varepsilon. \tag{10}$$

Since $\tilde{r} \in \mathrm{B}\left(SAS\right)$, there is no loss of generality in

assuming that $y_n \to y$ as $n \to \infty$ for some $y \in \Re$. Obviously it holds from Assumption A 2) that

$$\limsup_{n\to\infty} \bigcup \left(y_n \middle| x_n, a_n, \tilde{r} \right) \subset \bigcup \left(y \middle| x, a, \tilde{r} \right). \tag{11}$$

We show that

$$\liminf_{n\to\infty} \bigcup \left(y_n \middle| x_n, a_n, \tilde{r} \right) \supset \bigcup^0 \left(y \middle| x, a, \tilde{r} \right). \tag{12}$$

For any $z \in \bigcup^0 \left(y \middle| x, a, \tilde{r} \right), -\tilde{r}(x, a, z) < y$, so that there exists $\varepsilon_1, \varepsilon_2 > 0$ such that $-r(x, a, z) + \varepsilon_1 < y - \varepsilon_2$. Therefore, from Assumption A 2) and convergence assumptions there exists N for which $z \in \bigcup \left(y_n \middle| x_n, a_n \right)$ for $n \geq N$, which implies (12). Thus, by the general convergence theorem (cf. [17]) and (11) and (12), we have that

$$\begin{aligned} &\limsup_{n\to\infty} \mathrm{Q}\left(\bigcup \left(y_n \middle| x_n, a_n, \tilde{r} \right) \middle| x_n, a_n \right) \\ &\qquad \leq \mathrm{Q}\left(\limsup_{n\to\infty} \bigcup \left(y_n \middle| x_n, a_n, \tilde{r} \right) \middle| x, a \right) \\ &\qquad \leq \mathrm{Q}\left(\bigcup \left(y \middle| x, a, \tilde{r} \right) \middle| x, a \right), \end{aligned}$$

and

$$\begin{aligned} &\liminf_{n\to\infty} \mathrm{Q}\left(\bigcup \left(y_n \middle| x_n, a_n, \tilde{r} \right) \middle| x_n, a_n \right) \\ &\qquad \geq \mathrm{Q}\left(\liminf_{n\to\infty} \bigcup \left(y_n \middle| x_n, a_n, \tilde{r} \right) \middle| x, a \right) \\ &\qquad \geq \mathrm{Q}\left(\bigcup^0 \left(y \middle| x, a, \tilde{r} \right) \middle| x, a \right). \end{aligned}$$

By Assumption A 3), it holds that

$$\lim_{n\to\infty} \mathrm{Q}\left(\bigcup \left(y_n \middle| x_n, a_n, \tilde{r} \right) \middle| x_n, a_n \right) = \mathrm{Q}\left(\bigcup \left(y \middle| x, a, \tilde{r} \right) \middle| x, a \right).$$

Thus, together with (10), we get

$$\mathrm{Q}\left(\bigcup \left(y \middle| x, a, \tilde{r} \right) \middle| x, a \right) \geq \gamma$$

and $y \leq d + \varepsilon$, which shows that $\underline{D}$ is closed. Similarly, we can prove that

$$\overline{D} := \left\{ (x, a) \in SA \middle| D_{-\tilde{r}}^{-1}(\gamma | x, a) \geq d \right\}$$

is closed for and $d \in \Re$, *i.e.*, $D_{-\tilde{r}}^{-1}(\gamma | x, a)$ is upper semi-continuous in $(x, a) \in K$ This shows that $D_{-\tilde{r}}^{-1}(\gamma | x, a)$ is continuous in $(x, a) \in K$ as required. □

We can be in a position to state the main theorem in the discounted case.

Theorem 2.3. *Suppose that Assumption A holds. Then*,

1) The value function ρ_{DS} is given by

$$\rho_{DS}(\tilde{r}) = \int h_{DS}(\tilde{r} | x) \nu(\mathrm{d}x), \tag{13}$$

where $h_{DS}(\tilde{r} | \cdot) \in \mathrm{B}(S)$ is a unique solution to the optimality equation of the discounted case,

$$h_{DS}(\tilde{r} | x) = \min_{a \in A} \{ r(x, a) + \beta \int h_{DS}(\tilde{r} | y) \mathrm{Q}(\mathrm{d}y | x, a) \} \tag{14}$$

for $x \in S$.

2) The exists a measurable function $f^* : S \to A$ with $f^*(x) \in A(x)$ for each $x \in S$ such that $f^*(x)$ attains the minimum in (14) and the stationary policy f^* is discount risk-optimal.

Proof. By Lemma 2.2, $D_{-\tilde{r}}^{-1}(\gamma | x, a)$ is continuous in $(x, a) \in K$. Thus, from the definition (9) of $r(x, a)$ and Assumption A 4), we observe that $r(x, a)$ is continuous in $(x, a) \in K$ Thus, applying the theory of discounted MDPs (cf. Theorem 4.2.3. in [18]), the assertions of Theorem 2.3 follows. This completes the proof. □

2.3. The Average Case

In order to obtain the optimality equation for the average case, we assume that Assumption below holds, which guarantees the ergodicity of the process.

Assumption B. There exists a number $\alpha \in (0,1)$ such that

$$\sup_{x, x' \in S, a, a' \in A} \left\| \mathrm{Q}(\cdot | x, a) - \mathrm{Q}(\cdot | x', a') \right\| \leq 2\alpha, \tag{15}$$

where $\|\cdot\|$ denotes the variation norm for signed measures.

One of sufficient condition for Assumption B to hold, easily checked for applications, is as follows (cf. [19, 20]).

Assumption B' There exists a measure η on B_S with $\eta(S) > 0$ such that

$$\mathrm{Q}(\mathrm{D} | x, a) \geq \eta(\mathrm{D}) \quad \text{for all} \quad \mathrm{D} \in \mathrm{B}_S. \tag{16}$$

Theorem 2.4. *Suppose that Assumptions A and B hold. Then, there exists* $v \in \mathrm{B}(S)$ *such that*

$$\rho_{AV}(\tilde{r}) + v(x) = \min_{a \in A} \left\{ r(x, a) + \int v(y) \mathrm{Q}(\mathrm{d}y | x, a) \right\}. \tag{17}$$

Moreover, there is an average risk-optimal stationary policy f^* such that $f^*(x) \in A$ minimizes the right-hand side of (17).

Proof. We have already obtained that $r(x, a)$ is continuous in $(x, a) \in K$ So, applying the theory of average MDPs (cf. Corollary 3.6 in [19]), Theorem 2.4 follows, as required. □

3. An Application to Inventory Model

We consider the single-item model with a finite capacity $C < \infty$, in which the demands $\{\xi_t\}_{t=0}$ in successive periods are i.i.d. with the distribution function Φ on $\Re^+ = (0, \infty)$ which has a continuous density $\phi(x)$ w.r.t. the Lebesgue measure μ. The state space and

action space are $S = A = [0, C]$ and the set of admissible actions in state $x \in S$ is $A(x) = [0, C - x]$. The state X_t denotes the stock level at the beginning of period *t* and action Δ_t is the quantity ordered (and immediate supplied) at the beginning of period *t*. Putting the amount sold during period *t*, $Y_t = \min\{\xi_t, X_t + \Delta_t\}$, the system equation is given as follows.

$$X_{t+1} = X_t + \Delta_t - Y_t = [X_t + \Delta_t - \xi_t]^+ \quad (t = 0,1,2,\cdots). \tag{18}$$

The transition probability $Q(\cdot|x,a)$, for any Borel subset *B* of *S*, becomes

$$Q(B|x,a) = \int \mathbf{1}_B \{[x + a - y]^+\} \varphi(y) \mathrm{d}\mu. \tag{19}$$

Also, the immediate reward $\tilde{r} \in B(S \times A \times S)$ is given as

$$\tilde{r}(x,a,y) = p(x + a - y) - ca - h(x + a),$$

where $p > 0$ is the unit sale price, $c > 0$ the unit production cost and $h > 0$ unit holding cost. Several lemmas are needed for risk analysis. Let ξ be a random variable with a given demand distribution Φ and $Y = \min\{\xi, x\}$ for $x \in \Re$.

Lemma 3.1. *For* $\gamma \in (0,1)$, $CV@R_\gamma(Y)$ *is given as*

$$CV@R_\gamma(Y) = \begin{cases} CV@R_\gamma(\xi) & \text{if } 1-\gamma \le \overline{p}, \\ CV@R_\gamma(\xi) + \dfrac{1-\gamma-\overline{p}}{1-\gamma} & \text{if } 1-\gamma > \overline{p}, \end{cases} \tag{20}$$

where $\overline{p} = \Phi(x)$.

Proof. Recall that

$$CV@R_\gamma(Y) = -\frac{1}{1-\gamma}\int_0^{1-\gamma} F_Y^{-1}(p)\mathrm{d}p.$$

Since $F_Y^{-1}(p) = F_\xi^{-1}(p)$ if $p < \overline{p}$, $= x$ if $p \ge \overline{p}$, (20) follows obviously. □

In order to the equivalent MDPs, we specify the immediate reward $r \in B(S \times A \times S)$ by

$$\begin{aligned} & r(x,a) \\ &= CV@R_\gamma(\tilde{r}(x,a)|X_t = x, \Delta = a) \\ &= CV@R_\gamma(p\min\{\xi, x+a\} - ca - h(x+a)) \\ &= p \cdot CV@R_\gamma(\min\{\xi, x+a\}) + ca + h(x+a) \\ &= L(x+a) + ca, \end{aligned} \tag{21}$$

where $L(u) = p \cdot CV@R_\gamma(\min\{\xi, u\}) + hu, u \in \Re$ and the third equality follows from the monotonicity and homogeneous property of *CV@R*. The function *L* defined above is proved to be a convex function.

Lemma 3.2 *The following* 1) - 2) *hold.*

1) $\min\{a+b, c+d\} \ge \min\{a,c\} + \min\{b,d\}$.

2) The function $L(u)$ is convex.

Proof. The proof of 1) is easy, so omitted. Noting from 1) that $\min\{\xi, \lambda u_1 + (1-\lambda)u_2\} \ge \lambda \min\{\xi, u_1\} + (1-\lambda)\min\{\xi, u_2\}(u_1, u_2 \in \Re)$. For any $\lambda \in (0,1)$, we have that

$$\begin{aligned} & CV@R_\gamma(\min\{\xi, \lambda u_1 + (1-\lambda)u_2\}) \\ &= CV@R_\gamma(\min\{\lambda\xi + (1-\lambda)\xi, \lambda u_1 + (1-\lambda)u_2\}) \\ &\le CV@R_\gamma(\lambda \min\{\xi, u_1\} + (1-\lambda)\min\{\xi, u_2\}) \\ &\le \lambda \cdot CV@R_\gamma(\min\{\xi, u_1\}) + (1-\lambda) CV@R_\gamma(\min\{\xi, u_2\}). \end{aligned} \tag{22}$$

The second and the third inequalities follow from the monotonicity and the convexity of *CV@R*, respectively. This means that $L(u)$ is convex. □

To applying Theorems 2.3 and 2.4 to inventory problems, the following is needed.

Assumption C. It holds that $\delta := \int_c^\infty \phi(y)\mathrm{d}y > 0$.

We can state the main theorem.

Theorem 3.3. *Suppose that Assumption C holds. Then, for each of discounted or average case, there exists a constant level stationary policy* f^* *which is optimal, that is, the ordered amount* $f^*(x)$ *is*

$$f^*(x) = \begin{cases} x^* - x & \text{if } x < x^* \\ 0 & \text{if } x \ge x^* \end{cases} \tag{23}$$

for some $x^* \in \Re$, where the critical level x^* for each case is given from the corresponding optimality Equations (14) and (17).

Proof. First we verify that 1) - 4) of Assumption A are satisfied. A 1) – A 4) are clearly true by definitions. For any $v \in B([0,C])$, from (19) it holds that

$$\int v(y) Q(\mathrm{d}y|x,a) = \int v(y) \phi([x + a - y]^+)\mathrm{d}y,$$

which is continuous in

$$(x,a) \in K = \{(x,a) | 0 \le a \le C - x, x \in [0,C]\},$$

applying the dominated convergence theorem. We set $\eta(\mathrm{D}) = \eta \mathbf{1}_{\{0\}}(\mathrm{D})$. Then, assertion (16) in Assumption B holds. Thus, Theorems 2.3 and 2.4 are applicable. Since $r(x,a)$ is convex in a, using the result of Iglehant [21] (cf. [22]), it follows that the right-hand sides of the corresponding optimality equation (14) and (16) are convex in $a \in [0, C - x]$. So, it is easily shown that there exists a risk-optimal policy f^* of a constant level type (23) for each case. The proof is complete. □

4. Acknowledgements

This study was partly supported by "Development of

methodologies for risk trade-off analysis toward optimizing management of chemicals" funded by New Energy and Industrial Technology Development Organization (NEDO).

5. References

[1] H. M. Markowitz, "Portfolio Selection: Efficient Diversification of Investment," Wiley, New York, 1958.

[2] R. T. Rockafellar and S. Uryasev, "Optimization of Conditional Value-at-Risk," *Journal of Risk*, Vol. 2, No. 3, 2000, pp. 21-42.

[3] R. T. Rockafellar and S. Uryasev, "Conditional Value-at-Risk for General Loss Distributions," *Journal of Banking & Finance*, Vol. 26, No. 7, 2002, pp. 1443-1471. doi:10.1016/S0378-4266(02)00271-6

[4] P. Artzner, F. Delbaen, J. M. Eber and D. Heath, "Coherent Measure of Risk," *Mathematical Finance*, Vol. 9, 1999, pp. 203-227. doi:10.1111/1467-9965.00068

[5] A. Inoue, "On the Worst Conditional Expectation," *Journal on Applied Mathematics*, Vol. 286, No. 1, 2003, pp. 237-247.

[6] S. Kusuoka, "On Law Invariant Coherent Risk Measures," *Advances in Mathematical Economics*, Vol. 3, Springer, Tokyo, 2001, pp. 83-95.

[7] H. Föllmer and I. Penner, "Convex Measures of Risk and Trading Constraints," *Finance and Stochastics*, Vol. 6, No. 4, 2002, pp. 429-447. doi:10.1007/s007800200072

[8] H. Föllmer and I. Penner, "Convex Risk Measure and the Dynamics of Their Penalty Functions," *Statistics & Decision*, Vol. 24, 2006, pp. 61-96.

[9] J. Goto and Y. Takano, "Newsvendor Solutions via Conditional Value-at-Risk Minimization," *European Journal Operational Research*, Vol. 179, No. 1, 2007, pp. 80-96. doi:10.1016/j.ejor.2006.03.022

[10] A. Takeda, "Generalization Performance of ν-Support Vector Classifier Based on Conditional Value-at-Risk Minimization," *Neurocomputing*, Vol. 72, 2009, pp. 2351-2358.

[11] B. Kang and J. A. Filar, "Time Consistent Dynamic Risk Measures," *Mathematical Methods in Operations Research* 2005, Special Issue in Honor of Arice Hordijk 2005, pp. 1-19.

[12] Y. Ohtsubo and K. Toyonaga, "Optimal Policy for Minimizing Risk Models in Markov Decision Processes," *Journal of Mathematical Analysis and Applications*, Vol. 271, No. 1, 2002, pp. 66-81. doi:10.1016/S0022-247X(02)00097-5

[13] Y. Ohtsubo, "Optimal Threshold Probability in Discounted Markov Decision Processes with a Target Set," *Applied Mathematics and Computation*, Vol. 149, No. 2, 2004, pp. 519-532. doi:10.1016/S0096-3003(03)00158-9

[14] D. J. White, "Minimising a Threshold Probability in Discounted Markov Decision Processes," *Journal of Mathematical Analysis and Applications*, Vol. 173, No. 2, 1993, pp. 634-646. doi:10.1006/jmaa.1993.1093

[15] C. Wu and Y. Lin, "Minimizing Risk Models in Markov Decision Processes with Policies Depending on Target Values," *Journal of Mathematical Analysis and Applications*, Vol. 231, No. 1, 1999, pp. 47-67. doi:10.1006/jmaa.1998.6203

[16] A. P. Mundt, "Dynamic Risk Management with Markov Decision Processes," Universitätsverlag Karlsruhe, Karlsruhe, 2007.

[17] H. L. Royden, "Real Analysis," 2nd Edition, The Macmillan Company, New York, 1968.

[18] O. Hernández-Lerma and J. B. Lasserre, "Discrete-Time Markov Control Processes, Basic Optimality Criteria," Springer-Verlag, New York, 1995.

[19] O. Hernández-Lerma, "Adaptive Markov Control Processes," Springer-Verlag, New York, 1989.

[20] M. Kurano, "Markov Decision Processes with a Borel Measurable Cost Function: The Average Case," *Mathematics of Operations Research*, Vol. 11, No. 2, 1986, pp. 309-320.

[21] D. L. Iglehant, "Optimality of (*s*, *S*) Policies in the Infinite Horizon Dynamic Inventory Problem," *Management science*, Vol. 9, No. 2, 1963, pp. 259-267. doi:10.1287/mnsc.9.2.259

[22] S. M. Ross, "Applied Probability Models with Optimization Applications," Holden-Day, San Fransico, 1970.

