A note on risk evaluation in Markov decision processes

2010年7月9日　統計数理研究所　オープンハウス

影山 正幸　　リスク解析戦略研究センター　特任研究員

1 Introduction

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.

2 Preliminaries

Let I be a random income(or reward) variable on some probability space $(\Omega, \mathcal{B}, P)$, and $F_I(x)$ the distribution function of I, i.e., $F_I(x) = P(I \leq x)(x \in \Re)$. We define the inverse function $F_I^{-1}(p)(0 \leq p \leq 1)$ by $F_I^{-1}(p) = \inf\{x \in \Re | F_I(x) \geq p\}$. Then, the Conditional Value-at-Risk for a level $\gamma \in (0,1)$ of I, $CV@R_\gamma(I)$, is defined (cf. [2]) by

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

A Markov decision process is a controlled dynamic system defined by a six-tuple $\{S, A, \{A(x)|x \in A\}, \mathcal{Q}, \tilde{r}, \nu\}$, 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 $x \in S$, $\mathcal{Q} \in \mathcal{P}(S|SA)$ is the law of motion, $\tilde{r} \in \mathcal{B}(SAS)$ is an immediate reward function and $\nu \in \mathcal{P}(S)$ is an initial state distribution. 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 $(t \geq 0)$. 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.
(a) The discounted case$(0 < \beta < 1)$.

$$\rho_{DS}(\tilde{r}|\pi) := \frac{1}{1-\beta}\sum_{t=1}^{\infty}\beta^t E_\pi[CV@R_\gamma(\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t)|H_{t-1})]. \tag{2}$$

(b) The average case.

$$\rho_{AV}(\tilde{r}|\pi) := \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} E_\pi[CV@R_\gamma(\tilde{r}(X_{t-1}, \Delta_{t-1}, X_t)|H_{t-1})]. \tag{3}$$

Proposition([1]) For any $\pi \in \Pi$, ρ_{DS} and ρ_{AV} have the following (i)-(iv):
(i) (Monotonicity) If $\tilde{r}_1 \leq \tilde{r}_2$ with $\tilde{r}_1, \tilde{r}_2 \in \mathcal{B}(SAS)$, $\rho(\tilde{r}_1) \geq \rho(\tilde{r}_2)$.
(ii) (Translation invariance) For $\tilde{r} \in \mathcal{B}(SAS)$ and $c \in \Re = (-\infty, \infty)$, $\rho(\tilde{r} + c) = \rho(\tilde{r}) - c$.
(iii) (Homogeneity) For $\tilde{r} \in \mathcal{B}(SAS)$ and $\lambda > 0$, $\rho(\lambda\tilde{r}) = \lambda\rho(\tilde{r})$.
(iv) (Convexity) For $\tilde{r}_1, \tilde{r}_2 \in \mathcal{B}(SAS)$ and $0 \leq \lambda \leq 1$, $\rho(\lambda\tilde{r}_1 + (1-\lambda)\tilde{r}_2) \leq \lambda\rho(\tilde{r}_1) + (1-\lambda)\rho(\tilde{r}_2)$.

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

$$r(x,a) = D_{-\tilde{r}}^{-1}(\gamma|x,a) + \frac{1}{1-\gamma}\int\left[-\tilde{r}(x,a,y) - D_{\tilde{r}}^{-1}(\gamma|x,a)\right]^+\mathcal{Q}(dy|x,a) \tag{4}$$

for each $x \in S$ and $a \in A$.
Theorem 1([1]) It holds that, for any $\pi \in \Pi$,
(i) $\rho_{DS}(\tilde{r}|\pi) = \frac{1}{1-\beta}\sum_{t=0}^{\infty}\beta^t E^\pi[r(X_t, \Delta_t)]$,
(ii) $\rho_{AV}(\tilde{r}|\pi) = \limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} E^\pi[r(X_t, \Delta_t)]$.

3.1 The discounted case

Assumption A The following (i)-(iv) holds:
(i) A is compact and $A(x)$ is closed for each $x \in A$.
(ii) $\tilde{r}(x,a,y) \in \mathcal{B}(SAS)$ is continuous in $(x,a,y) \in SAS$.
(iii) $\mathcal{Q}(\partial\cup(y|x,a,\tilde{r})|x,a) = 0$ for each $(x,a) \in K$ and $y \in \Re$, where $\partial\cup(y|x,a,\tilde{r}) = \{z \in S| - \tilde{r}(x,a,z) = y\}$.
(iv) $\mathcal{Q}(\cdot|x,a)$ is strongly continuous in $(x,a) \in K$, i.e., for any $v \in \mathcal{B}(S)$, $\int v(y)\mathcal{Q}(dy|x,a)$ is continuous in $(x,a) \in K$.
Theorem 2([1]) Suppose that Assumption A holds. Then,
(i) The value function ρ_{DS} is given by

$$\rho_{DS}(\tilde{r}) = \int h_{DS}(\tilde{r}|x)\nu(dx), \tag{5}$$

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

$$h_{DS}(\tilde{r}|x) = \min_{a\in A}\Big\{r(x,a) + \beta\int h_{DS}(\tilde{r}|y)\mathcal{Q}(dy|x,a)\Big\} \text{ for } x \in S. \tag{6}$$

(ii) 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 (6) and the stationary policy f^* is discount risk-optimal.

3.2 The average case

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

$$\sup_{x,x'\in S, a,a'\in A}\left\|\mathcal{Q}(\cdot|x,a) - \mathcal{Q}(\cdot|x',a')\right\| \leq 2\alpha, \tag{7}$$

where $\|\cdot\|$ denotes the variation norm for signed measures.
Theorem 3([1]) Suppose that Assumptions A and B hold. Then, there exists $v \in \mathcal{B}(S)$ such that

$$\rho_{AV}(\tilde{r}) + v(x) = \min_{a\in A}\Big\{r(x,a) + \int v(y)\mathcal{Q}(dy|x,a)\Big\}. \tag{8}$$

Moreover, there is an average risk-optimal stationary policy f^* such that $f^*(x) \in A$ minimizes the right-hand side of (8).

4 An application to inventory model

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.

$$X_{t+1} = X_t + \Delta_t - Y_t = [X_t + \Delta_t - \xi_t]^+ \quad (t = 0, 1, 2, \ldots). \tag{9}$$

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

$$\mathcal{Q}(B|x,a) = \int \mathbf{1}_B\{[x+a-y]^+\}\phi(y)d\mu. \tag{10}$$

Also, the immediate reward $\tilde{r} \in \mathcal{B}(S \times A \times S)$ 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.
Assumption C It holds that $\delta := \int_c^\infty \phi(y)dy > 0$.
Theorem 4([1]) Suppose that Assumption C holds. Then, for each of discounted or average case, there exists an optimal stationary policy f^*, whose ordered amount $f^*(x)$ is

$$f^*(x) = \begin{cases} x^* - x & \text{if } x < x^* \\ 0 & \text{if } x \geq x^* \end{cases} \tag{11}$$

for some $x^* \in \Re$, where the critical number x^* for each case is given from the corresponding optimality equations (6) and (8).

References

[1] Kageyama M., Fujii T., Kanefuji K. and Tsubaki H., Conditional value-at-risk for random immediate reward variables in Markov decision processes, submitted.

[2] Rockafellar RT. and Uryasev S., Optimization of conditional value-at-risk, *Journal of Risk*, **2**, 21-42 (2000)