Risk-Sensitive Optimality Criteria in Markov Decision Processes

Karel Sladký

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Praha, Czech Republic
sladky@utia.cas.cz

1 Introduction and Notation

The usual optimization criteria for Markov decision processes (e.g. total discounted reward or mean reward) can be quite insufficient to fully capture the various aspects for a decision maker. It may be preferable to select more sophisticated criteria that also reflect variability-risk features of the problem. To this end we focus attention on risk-sensitive optimality criteria (i.e. the case when expectation of the stream of rewards generated by the Markov processes evaluated by an exponential utility function is considered) and their connections with mean-variance optimality (i.e. the case when a suitable combination of the expected total reward and its variance, usually considered per transition, is selected as a reasonable optimality criterion). The research of risk-sensitive optimality criteria in Markov decision processes was initiated in the seminal paper by Howard and Matheson [6] and followed by many other researchers (see e.g. [1, 2, 3, 5, 4, 8, 9, 14]). In this note we consider a Markov decision chain $X = \{X_n, n = 0,1,\ldots\}$ with finite state space $\mathcal{I} = \{1,2,\ldots,N\}$ and a finite set $\mathcal{A}_i = \{1,2,\ldots,K_i\}$ of possible decisions (actions) in state $i \in \mathcal{I}$. Supposing that in state $i \in \mathcal{I}$ action $k \in \mathcal{A}_i$ is selected, then state j is reached in the next transition with a given probability p_{ij}^k and one-stage transition reward r_{ij} will be accrued to such transition.

Suppose that the stream of transition rewards is evaluated by an exponential utility function, say $u^\gamma(\cdot)$, i.e. a utility function with constant risk sensitivity $\gamma \in \mathbb{R}$. Then the utility assigned to the (random) reward ξ is given by

$$u^\gamma(\xi) := \begin{cases} (\text{sign } \gamma)\, \exp(\gamma\xi), & \text{if } \gamma \neq 0 \\ \xi & \text{for } \gamma = 0. \end{cases} \tag{1}$$

Obviously $u^\gamma(\cdot)$ is continuous and strictly increasing, and convex (resp. concave) for $\gamma > 0$ (resp. $\gamma < 0$). If ξ is a (bounded) random variable then for the corresponding certainty equivalent of the (random) variable ξ, say $Z(\xi)$,

in virtue of the condition $u^\gamma(Z(\xi)) = \mathsf{E}\,[(\text{sign }\gamma)\exp(\gamma\xi)]$ (E is reserved for expectation), we immediately get

$$Z(\xi) = \begin{cases} \frac{1}{\gamma}\ln\{\mathsf{E}\,[\exp(\gamma\xi)]\}, & \text{if } \gamma \neq 0 \\ \mathsf{E}\,[\xi] & \text{for } \gamma = 0. \end{cases} \tag{2}$$

Observe that if ξ is constant then $Z(\xi) = \xi$, if ξ is nonconstant then by Jensen's inequality

$$\begin{aligned} Z(\xi) &> \mathsf{E}\,\xi \quad (\text{if } \gamma > 0 \text{ and the decision maker is risk averse}) \\ Z(\xi) &< \mathsf{E}\,\xi \quad (\text{if } \gamma < 0 \text{ and the decision maker is risk seeking}) \\ Z(\xi) &= \mathsf{E}\,\xi \quad (\text{if } \gamma = 0 \text{ and the decision maker is risk neutral}) \end{aligned}$$

A (Markovian) policy, say π, controlling the chain is a rule how to select actions in each state. We write $\pi = (f^0, f^1, \ldots)$ where $f^n \in \mathcal{A} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$ for every $n = 0,1,2,\ldots$ and $f_i^n \in \mathcal{A}_i$ is the decision at the nth transition when the chain X is in state i. A policy which takes at all times the same decision rule, i.e. selects actions only with respect to the current state and hence is fully identified by some decision vector f whose ith element $f_i \in \mathcal{A}_i$, is called stationary. Stationary policy $\pi \sim (f)$ then completely identifies the transition probability matrix $\boldsymbol{P}(f)$ along with the one-stage expected reward vector $\boldsymbol{r}(f)$. Observe that the ith row of $\boldsymbol{P}(f)$, denoted $\boldsymbol{p}_i(f)$, has elements $p_{i1}^{f_i}, \ldots, p_{iN}^{f_i}$ and that $\boldsymbol{P}^*(f) = \lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1}[\boldsymbol{P}(f)]^k$ exists. Similarly the ith element of $\boldsymbol{r}(f)$ denotes the one-stage expected value $r_i^{f_i} = \sum_{k=0}^{n-1} p_{ij}^{f_i} r_{ij}$.

Let $\xi_n = \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}$ be the stream of transition rewards received in the n next transitions of the considered Markov chain X, and similarly let $\xi^{(m,n)}$ be reserved for the total (random) reward obtained from the mth up to the nth transition (obviously, $\xi_n = r_{X_0,X_1} + \xi^{(1,n)}$). In case that $\gamma \neq 0$ then $u^\gamma(\xi_n) := (\text{sign}\,\gamma)\mathrm{e}^{\gamma\sum_{k=0}^{n-1} r_{X_k,X_{k+1}}}$ is the (random) utility assigned to ξ_n, and $Z(\xi_n) = \frac{1}{\gamma}\ln\{\mathsf{E}\,[\mathrm{e}^{\gamma\sum_{k=0}^{n-1} r_{X_k,X_{k+1}}}]\}$ is its certainty equivalent. Obviously, if $\gamma = 0$ then $u^\gamma(\xi_n) = \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}$ and $Z(\xi_n) = \mathsf{E}\,[\sum_{k=0}^{n-1} r_{X_k,X_{k+1}}]$.

Supposing that the chain starts in state $X_0 = i$ and policy $\pi = (f^n)$ is followed, then if $\gamma \neq 0$ for the expected utility in the n next transitions and the corresponding certainty equivalent we have (E_i^π denotes expectation if policy π is followed and $X_0 = i$)

$$U_i^\pi(\gamma,0,n) := \mathsf{E}_i^\pi[u^\gamma(\xi_n)] = (\text{sign }\gamma)\mathsf{E}_i^\pi[\exp(\gamma\sum_{k=0}^{n-1} r_{X_k,X_{k+1}})] \tag{3}$$

$$Z_i^\pi(\gamma,0,n) := \frac{1}{\gamma}\ln\{\mathsf{E}_i^\pi[\exp(\gamma\sum_{k=0}^{n-1} r_{X_k,X_{k+1}})]\}. \tag{4}$$

In what follows we shall often abbreviate $U_i^\pi(\gamma,0,n)$ (resp. $Z_i^\pi(\gamma,0,n)$) by $U_i^\pi(\gamma,n)$ (resp. $Z_i^\pi(\gamma,n)$). Similarly $\boldsymbol{U}^\pi(\gamma,n)$ (resp. $\boldsymbol{Z}^\pi(\gamma,n)$) is reserved for