RISK-SENSITIVE AND RISK NEUTRAL OPTIMALITY IN MARKOV DECISION CHAINS; A UNIFIED APPROACH

Karel Sladký
Institute of Information Theory and Automation of the AS CR

Abstract: In this note we consider Markov decision chains with finite state space and compact actions spaces where the stream of rewards generated by the Markov processes is evaluated by an exponential utility function (so-called risk-sensitive model) with a given risk sensitivity coefficient. If the risk sensitivity coefficient equals zero (risk-neutral case) we arrive at a standard Markov decision chain. Necessary and sufficient optimality conditions along with equations for average optimal policies both for risk-neutral and risk-sensitive models will be presented and connections and similarity between these approaches will be discussed.

Keywords: discrete-time Markov decision chains, exponential utility functions, risk sensitivity coefficient, connections between risk-sensitive and risk-neutral models

JEL Classification: C44
AMS Classification: 90C40

1 Introduction and Notation

In this note, we consider Markov decision processes with finite state and compact action spaces where the stream of rewards generated by the Markov processes is evaluated by an exponential utility function (so-called risk-sensitive model) with a given risk sensitivity coefficient, and slightly extend some of the results reported in [1,2,9–12]. To this end, let us consider an exponential utility function, say $\bar{u}^\gamma(\cdot)$, i.e. a separable utility function with constant risk sensitivity $\gamma \in R$, where the utility assigned to the (random) outcome ξ is given by

$$\bar{u}^\gamma(\xi) := \begin{cases} (\text{sign } \gamma)\, \exp(\gamma\xi), & \text{if } \gamma \neq 0, \\ \xi \quad \text{for } \gamma = 0 \quad \text{(the risk-neutral case).} \end{cases} \tag{1}$$

For what follows let $u^\gamma(\xi) := \exp(\gamma\xi)$, hence $\bar{u}^\gamma(\xi) = (\text{sign } \gamma)u^\gamma(\xi)$. Then for the corresponding certainty equivalent, say $Z^\gamma(\xi)$, since $\bar{u}^\gamma(Z^\gamma(\xi)) = \mathrm{E}[\bar{u}^\gamma(\xi)]$ (E is reserved for expectation), we immediately get

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

In what follows, we consider at discrete time points Markov decision process $X = \{X_n, n = 0, 1, \ldots\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$, and compact set $\mathcal{A}_i = [0, K_i]$ of possible decisions (actions) in state $i \in \mathcal{I}$. Supposing that in state $i \in \mathcal{I}$ action $a \in \mathcal{A}_i$ is chosen, then state j is reached in the next transition with a given probability $p_{ij}(a)$ and one-stage transition reward r_{ij} will be accrued to such transition.

A (Markovian) policy controlling the decision process is given by a sequence of decisions at every time point. In particular, policy controlling the process, $\pi = (f^0, f^1, \ldots)$, is identified by a sequence of decision vectors $\{f^n, n = 0, 1, \ldots\}$ where $f^n \in \mathcal{F} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$ for every

$n = 0, 1, 2, \ldots$, and $f_i^n \in \mathcal{A}_i$ is the decision (or action) taken at the nth transition if the chain X is in state i. Policy π which selects at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary. We shall assume that the stream of transition rewards generated by the considered Markov decision process is evaluated by an exponential utility function (1). To this end, let $\xi_{X_0}^n(\pi) = \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}$ be the (random) total reward received in the n next transitions of the considered Markov chain X if policy $\pi = (f^n)$ is followed and the chain starts in state X_0. Supposing that $X_0 = i$, on taking expectation we have (E_i^π denotes the expectation if $X_0 = i$ and policy $\pi = (f^n)$ is followed)

$$\bar{U}_i^\gamma(\pi, n) := \mathrm{E}_i^\pi(\bar{u}^\gamma(\xi^n)) = (\text{sign } \gamma)\mathrm{E}_i^\pi \mathrm{e}^{\gamma \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}} \quad \text{if} \quad \gamma \neq 0 \tag{3}$$

$$V_i(\pi, n) := \mathrm{E}_i^\pi(\bar{u}(\xi^n)) = \mathrm{E}_i^\pi \sum_{k=0}^{n-1} r_{X_k,X_{k+1}} \quad \text{if} \quad \gamma = 0, \quad \text{the risk neutral case.} \tag{4}$$

2 Risk-neutral Case: Optimality Equations

To begin with, (cf. [6, 8]) first observe that if the discrepancy function

$$\bar{\varphi}_{ij}(w, \bar{g}) := r_{ij} - \bar{g} + w_j - w_i, \quad \text{for arbitrary} \quad \bar{g}, w_i \in \mathbb{R}, i, j \in \mathcal{I} \tag{5}$$

then by (4)

$$V_i(\pi, n) = \bar{g} + w_i + \sum_{j \in \mathcal{I}} p_{ij}(f_i^0)\{\bar{\varphi}_{ij}(w, \bar{g}) + V_j(\pi^1, n-1) - w_j\} \tag{6}$$

$$= n\bar{g} + w_i + \mathrm{E}_i^\pi \sum_{k=0}^{n-1} \bar{\varphi}_{X_k,X_{k+1}}(w, \bar{g}) - \mathrm{E}_i^\pi w_{X_n} \tag{7}$$

For what follows we introduce matrix notation. We denote by $P(f) = [p_{ij}(f_i)]$ the $N \times N$ transition probability matrix of the chain X. Recall that the limiting matrix $P^*(f) = \lim_{m\to\infty} m^{-1} \sum_{n=0}^{m-1} P^n(f)$ exists, in particular, if $P(f)$ is *unichain* (i.e. $P(f)$ contains a single class of recurrent states) the rows of $P^*(f)$, denoted $p^*(f)$, are identical.

Obviously, $r_i(f_i) := \sum_{j=1}^N p_{ij}(f_i) r_{ij}$ (resp. $\varphi_i(f_i, w, \bar{g}) := \sum_{j=1}^N p_{ij}(f_i)\bar{\varphi}_{ij}(w, \bar{g})$) is the expected one-stage reward (resp. expected discrepancy) obtained in state $i \in \mathcal{I}$, and $r(f)$ (resp. $\varphi(f, w, \bar{g})$) denotes the corresponding N-dimensional column vector of one-stage rewards (resp. expected discrepancies). Then $[P(f)]^n \cdot r$ (resp. $[P(f)]^n \cdot \varphi(f, w, \bar{g})$) is the (column) vector of expected rewards (resp. expected discrepancies) accrued after n transitions; its ith entry denotes expectation of the reward (resp. discrepancy) obtained at time point n if the process X starts in state i.

Similarly, the vector of total expected rewards earned up to the n-th transition

$$V(\pi, n) := \sum_{k=0}^{n-1} \prod_{j=0}^{k-1} P(f^j) r(f^k) = ng + w + \sum_{k=0}^{n-1} \prod_{j=0}^{k-1} P(f^j)\varphi(f^k, w, \bar{g}) - \prod_{j=0}^{k-1} P(f^j) w \tag{8}$$

and its i-th element $V_i(\pi, n)$ is the total expected reward if the process starts in state i. Observe that for $n \to \infty$ elements of $V(\pi, n)$ can be typically infinite. Moreover, following stationary policy $\pi \sim (f)$ for n tending to infinity there exist vector of average expected rewards, denoted $g(f)$ (with elements $g_i(f)$) where $g(f) = \lim_{n\to\infty} \frac{1}{n} V(f, n) = P^*(\pi) r(f)$.

Assumption A. There exists state $i_0 \in \mathcal{I}$ that is accessible from any state $i \in \mathcal{I}$ for every $f \in \mathcal{F}$, i.e. for every $f \in \mathcal{F}$ the transition probability matrix $P(f)$ is *unichain*.

The following facts are well-known to workers in stochastic dynamic programming (see e.g. [4, 7]).

If Assumption A holds there exists decision vector $f^* \in \mathcal{F}$ (resp. $\hat{f} \in \mathcal{F}$) along with (column) vectors $w^* = w(f^*)$, $\hat{w} = w(\hat{f})$ with elements w_i^*, $\hat{w}_i$ respectively, and $g^* = g(f^*)$ (resp. $\hat{g} = g(\hat{f})$) (constant vector with elements $\bar{g}(f) = p^*(f)r(f)$) being the solution of the (nonlinear) equation (I denotes the identity matrix)

$$\max_{f\in\mathcal{F}} \left[r(f) - g^* + (P(f) - I)\cdot w^*\right] = 0, \qquad \min_{f\in\mathcal{F}} \left[r(f) - \hat{g} + (P(f) - I)\cdot \hat{w}\right] = 0 \tag{9}$$

where $w(f)$ for $f = f^*, \hat{f}$ is unique up to an additive constant, and unique under the additional normalizing condition $P^*(f)\ w(f) = 0$. Then for

$$\varphi(f, f^*) := r(f) - g(f^*) + (P(f) - I)\cdot w(f^*), \qquad \varphi(f, \hat{f}) := r(f) - g(\hat{f}) + (P(f) - I)\cdot w(\hat{f}) \tag{10}$$

we have $\varphi(f, f^*) \le 0$, $\varphi(f, \hat{f}) \ge 0$ with $\varphi(f, f^*) = \varphi(\hat{f}, \hat{f}) = 0$.

In particular, by (9)–(10) for every $i \in \mathcal{I}$ we can write

$$\varphi_i(f, f^*) = r_i(f_i) - \bar{g}^* + \sum_{j\in\mathcal{I}} p_{ij}(f_i)w_j^* - w_i^* \le 0, \quad \varphi_i(f, \hat{f}) = r_i(f_i) - \hat{\bar{g}} + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\hat{w}_j - \hat{w}_i \ge 0.$$

3 Risk-Sensitive Models: Optimality Equations

For the risk-sensitive models, let $U_i^\gamma(\pi, n) := \mathrm{E}_i^\pi(u^\gamma(\xi^n))$ and hence $Z_i^\gamma(\pi, n) = \frac{1}{\gamma}\ln U_i^\gamma(\pi, n)$ be the corresponding certainty equivalent. In analogy to (6), (7) for expectation of the utility function we get by (5) for arbitrary $\bar{g}$, $w_i \in \mathbb{R}$, $i, j \in \mathcal{I}$

$$\begin{aligned} U_i^\gamma(\pi, n) &= \mathrm{e}^{\gamma(\bar{g}+w_i)} \sum_{j\in\mathcal{I}} p_{ij}(f_i^0) e^{\gamma\{\bar{\varphi}_{ij}(w,\bar{g}) - w_j\}} \cdot U_j^\gamma(\pi^1, n-1) && (11)\\ &= \mathrm{e}^{\gamma(2\bar{g}+w_i)} \sum_{j\in\mathcal{I}}\sum_{k\in\mathcal{I}} p_{ij}(f_i^0) e^{\gamma\{\bar{\varphi}_{ij}(w,\bar{g}) - w_j\}} \cdot \mathrm{e}^{\gamma w_j} p_{jk}(f_j^1) e^{\gamma\{\bar{\varphi}_{jk}(w,\bar{g}) - w_k\}} \cdot U_k^\gamma(\pi^2, n-2) \\ &\ \ \vdots \\ &= \mathrm{e}^{\gamma(n\bar{g}+w_i)} \mathrm{E}_i^\pi e^{\gamma\{\sum_{k=0}^{n-1} \bar{\varphi}_{X_k, X_{k+1}}(w,\bar{g}) - w_{X_n}\}}. && (12) \end{aligned}$$

In particular, for stationary policy $\pi \sim (f)$ assigning numbers $g(f)$, $w_i(f)$ by (5) we have

$$\bar{\varphi}_{ij}(w(f), \bar{g}(f)) := r_{ij} - \bar{g}(f) + w_j(f) - w_i(f) \tag{13}$$

and (11),(12) take the form

$$\begin{aligned} U_i^\gamma(f, n) &= \mathrm{e}^{\gamma(\bar{g}(f)+w_i(f))} \sum_{j\in\mathcal{I}} p_{ij}(f_i) e^{\gamma\{\bar{\varphi}_{ij}(w(f),\bar{g}(f)) - w_j(f)\}} \cdot U_j^\gamma(f, n-1) \\ &= \mathrm{e}^{\gamma(n\bar{g}(f)+w_i(f))} \mathrm{E}_i^\pi e^{\gamma\{\sum_{k=0}^{n-1} \bar{\varphi}_{X_k, X_{k+1}}(w(f),\bar{g}(f)) - w_{X_n}(f)\}}. \end{aligned}$$

In what follows we show that under certain assumptions there exist $w_i(f)$'s, $g(f)$ such that

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)\mathrm{e}^{\gamma r_{ij}} \cdot \mathrm{e}^{\gamma w_j(f)} = \mathrm{e}^{\gamma\bar{g}(f)} \cdot \mathrm{e}^{\gamma w_i(f)}, \quad \text{for} \quad i \in \mathcal{I}. \tag{14}$$

Now let $\rho(f) := \mathrm{e}^{\gamma g(f)}$, $z_i(f) := \mathrm{e}^{\gamma w_i(f)}$, $q_{ij}(f_i) := p_{ij}(f_i)\mathrm{e}^{\gamma r_{ij}}$ and introduce the following matrix notation $U^\gamma(\pi, n) = [U_i^\gamma(\pi, n)]$, $z(f) = [z_i(f)]$ ... N-column vectors, $Q(f) = [q_{ij}(f_i)]$ ... $N \times N$ nonnegative matrix.

Then by (14) for stationary policy $\pi \sim (f)$ we immediately have $\rho(f)z(f) = Q(f)z(f)$. Since $Q(f)$ is a nonnegative matrix by the well-known Perron-Frobenius theorem $\rho(f)$ equals the spectral radius of $Q(f)$ and $z(f)$ can be selected nonnegative. Moreover, if $P(f)$ is irreducible then $Q(f)$ is irreducible, and $z(f)$ can be selected strictly positive (cf. [3]). Finally observe that and if $P(f)$ is unichain then $z(f)$ can be selected strictly positive if the risk sensitivity coefficient γ is sufficiently close to zero.

In (14) attention was focused only on a fixed stationary policy $\pi \sim (f)$. The above facts can be extended to all admissible policies under the following

Assumption B. There exists state $i_0 \in \mathcal{I}$ that for every $f \in \mathcal{F}$is accessible from any state $i \in \mathcal{I}$, i.e. for every $f \in \mathcal{F}$ the transition probability matrix $P(f)$ is unichain. Furthermore, if for some $f \in \mathcal{F}$ the matrices $P(f)$ and also $Q(f)$ are reducible then state i_0 belongs to the basic class of $Q(f)$ that is unique.

If Assumption B holds we can show existence of numbers w_i^* $(i \in \mathcal{I}), g^*$, and some $f^* \in \mathcal{F}$ such that for all $i \in \mathcal{I}$

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)e^{\gamma\{r_{ij}+w_j^*\}} \le \sum_{j\in\mathcal{I}} p_{ij}(f_i^*)e^{\gamma\{r_{ij}+w_j^*\}} = \mathrm{e}^{\gamma[g^*+w_i^*]} \tag{15}$$

or equivalently

$$\sum_{j\in\mathcal{I}} q_{ij}(f_i)z_j(f^*) \le \sum_{j\in\mathcal{I}} q_{ij}(f_i^*)z_j(f^*) = \rho(f^*)z_i(f^*). \tag{16}$$

Moreover, if Assumption B is fulfilled there also exist $\hat{w}_i$ $(i \in \mathcal{I}), \hat{g}$, and some $\hat{f} \in \mathcal{F}$ such that for all $i \in \mathcal{I}$

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)e^{\gamma\{r_{ij}+\hat{w}_j\}} \ge \sum_{j\in\mathcal{I}} p_{ij}(\hat{f}_i)e^{\gamma\{r_{ij}+\hat{w}_j\}} = \mathrm{e}^{\gamma[\hat{g}+\hat{w}_i]} \tag{17}$$

or equivalently

$$\sum_{j\in\mathcal{I}} q_{ij}(f_i)z_j(\hat{f}) \ge \sum_{j\in\mathcal{I}} q_{ij}(\hat{f}_i)z_j(\hat{f}) = \rho(\hat{f})z_i(\hat{f}). \tag{18}$$

Observe that by (17), (18) it holds for any $f \in \mathcal{F}$

$$Q(f)z(f^*) \le Q(f^*)z(f^*) = \rho(f^*)z(f^*), \qquad Q(f)z(\hat{f}) \ge Q(\hat{f})z(\hat{f}) = \rho(\hat{f})z(\hat{f}). \tag{19}$$

Theorem. If Assumption B holds there exists decision vector $f^* \in \mathcal{F}$ (resp. $\hat{f} \in \mathcal{F}$) along with column vector $z(f^*)$ (resp. $z(\hat{f})$) and a positive number $\rho(f^*)$, along with $g(f^*) = \ln \rho(f^*)$, (resp. $\rho(\hat{f})$, along with $g(\hat{f}) = \ln \rho(\hat{f})$) such that for any $f \in \mathcal{F}$ $\rho(\hat{f}) \le \rho(f) \le \rho(f^*)$ and also $g(\hat{f}) \le g(f) \le g(f^*)$.

The proof (by policy iterations) based on ideas in [5] can be found in [12].

4 Necessary and Sufficient Optimality Conditions

4.1 Risk-neutral case

To begin with, from Eq.(8) considered for decision vector f^* maximizing the average reward with $g = g^*$, $w = w^*$ we immediately have for policy $\pi = (f^n)$

$$V(\pi, n) := \sum_{k=0}^{n-1} \prod_{j=0}^{k-1} P(f^j) r(f^k) = ng^* + w^* + \sum_{k=0}^{n-1} \prod_{j=0}^{k-1} P(f^j)\varphi(f^k, f^*) - \prod_{j=0}^{k-1} P(f^j) w^*. \tag{20}$$

Hence for stationary policy $\pi^* \sim f^*$ maximizing average reward we immediately get

$$V(\pi^*, n) = ng^* + w^* - \prod_{j=0}^{k-1} P(f^j) w^* \tag{21}$$

and (nonstationary) policy $\pi = (f^n)$ maximizes long run average reward if and only if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \prod_{j=0}^{k-1} P(f^j)\varphi(f^k, f^*) = 0. \tag{22}$$

4.2 Risk-sensitive case

From Eq.(12) considered for decision vector f^* fulfilling conditions (17), (18) we immediately have for policy $\pi = (f^n)$

$$\begin{aligned} U_i^\gamma(\pi, n) &= \mathrm{e}^{\gamma(g^* + w_i^*)} \sum_{j\in\mathcal{I}} p_{ij}(f_i^0) e^{\gamma\{\bar{\varphi}_{ij}(w^*, g^*) - w_j^*\}} \cdot U_j^\gamma(\pi^1, n-1) && (23) \\ &= \mathrm{e}^{\gamma(ng^* + w_i^*)} \mathrm{E}_i^\pi e^{\gamma\{\sum_{k=0}^{n-1} \bar{\varphi}_{X_k, X_{k+1}}(w^*, g^*) - w_{X_n}^*\}}. && (24) \end{aligned}$$

Hence for stationary policy $\pi^* \sim (f^*)$ with f^* fulfilling conditions (17), (18) we immediately get

$$U_i^\gamma(f^*, n) = \mathrm{e}^{\gamma(ng^* + w_i^*)} \mathrm{E}_i^\pi e^{\{-\gamma w_{X_n}^*\}}. \tag{25}$$

Since the state space $\mathcal{I}$ is finite, there exists number $K > 0$ such that $|w_i^*| \le K$ for each $i \in \mathcal{I}$. Hence by (2), (24),(25) we immediately conclude that

$$U_i^\gamma(\pi, n) \le \mathrm{e}^{\gamma(ng^* + w_i^*)} \cdot \mathrm{e}^{|\gamma| K}, \quad Z_i^\gamma(\pi, n) = \frac{1}{\gamma} \ln U_i^\gamma(\pi, n) \tag{26}$$

In virtue of (17), (18), (19) from (26) we can conclude that for stationary policy $\pi \sim (f^*)$ or $\pi \sim (\hat{f})$ and arbitrary policy $\pi = (f^n)$

$$\lim_{n\to\infty} \frac{1}{n} Z_i^\gamma(\pi, n) = g^* \qquad \text{if and only if} \qquad \lim_{n\to\infty} \frac{1}{n} \ln \left\{ \mathrm{E}_i^\pi \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \bar{\varphi}_{X_k, X_{k+1}}(w^*, g^*)} \right\} = 0 \tag{27}$$

$$\lim_{n\to\infty} \frac{1}{n} Z_i^\gamma(\pi, n) = \hat{g} \qquad \text{if and only if} \qquad \lim_{n\to\infty} \frac{1}{n} \ln \left\{ \mathrm{E}_i^\pi \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \bar{\varphi}_{X_k, X_{k+1}}(\hat{w}, \hat{g})} \right\} = 0. \tag{28}$$

Acknowledgements: This research was supported by the Czech Science Foundation under Grants P402/11/0150 and P402/10/0956.

References

[1] Cavazos-Cadena R.: *Solution to the risk-sensitive average cost optimality equation in a class of Markov decision processes with finite state space.* Mathematical Methods of Operations Research **57** (2003), 253–285.

[2] Cavazos-Cadena R. and Montes-de-Oca R.: *The value iteration algorithm in risk-sensitive average Markov decision chains with finite state space.* Mathematics of Operations Research **28** (2003), 752–756.

[3] Gantmakher F.R.: *The Theory of Matrices.* Chelsea, London 1959.

[4] Howard R.A.: *Dynamic Programming and Markov Processes.* MIT Press, Cambridge, Mass. 1960.

[5] Howard R.A. and Matheson J.: *Risk-sensitive Markov decision processes.* Management Science **23** (1972), 356–369.

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

[7] Puterman M.L.: *Markov Decision Processes – Discrete Stochastic Dynamic Programming.* Wiley, New York 1994.

[8] Sladký K.: *On the Set of Optimal Controls for Markov Chains with Rewards.* Kybernetika **10** (1974), 526–547.

[9] Sladký K.: *On dynamic programming recursions for multiplicative Markov decision chains.* Mathematical Programming Study **6** (1976), 216–226.

[10] Sladký K.: *Bounds on discrete dynamic programming recursions I.* Kybernetika **16** (1980), 526–547.

[11] Sladký K.: *Risk sensitive discrete- and continuous-time Markov reward processes.* In: Proceedings of the International Scientific Conference Quantitative Methods in Economics (Multiple Criteria Decision Making XIV), M. Reiff, ed., University of Economics, Bratislava 2008, pp. 272–281.

[12] Sladký K.: *Growth rates and average optimality in risk-sensitive Markov decision chains.* Kybernetika **44** (2008), 205–226.

KAREL SLADKÝ
Department of Econometrics
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic
e-mail: sladky@utia.cas.cz