Risk-Sensitive and Average Optimality in Markov Decision Processes

Karel Sladký[1]

Abstract. This contribution is devoted to the risk-sensitive optimality criteria in finite state Markov Decision Processes. At first, we rederive necessary and sufficient conditions for average optimality of (classical) risk-neutral unichain models. This approach is then extended to the risk-sensitive case, i.e., when expectation of the stream of one-stage costs (or rewards) generated by a Markov chain is evaluated by an exponential utility function. We restrict ourselves on irreducible or unichain Markov models where risk-sensitive average optimality is independent of the starting state. As we show this problem is closely related to solution of (nonlinear) Poissonian equations and their connections with nonnegative matrices.

Keywords: dynamic programming, stochastic models, risk analysis and management.

JEL classification: C44, C61, C63
AMS classification: 90C40, 60J10, 93E20

1 Notation and Preliminaries

In this note, we consider unichain Markov decision processes with finite state space and compact action spaces where the stream of costs generated by the Markov processes is evaluated by an exponential utility function (so-called risk-sensitive models) with a given risk sensitivity coefficient.

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 $\gamma \in \mathbb{R}$. For $\gamma > 0$ (risk averse case) $\bar{u}^\gamma(\cdot)$ is convex, if $\gamma < 0$ (risk seeking case) $\bar{u}^\gamma(\cdot)$ is concave. Finally if $\gamma = 0$ (risk neutral case) $\bar{u}^\gamma(\cdot)$ is linear. Observe that exponential utility function $\bar{u}^\gamma(\cdot)$ is separable and multiplicative if the risk sensitivity $\gamma \neq 0$ and additive for $\gamma = 0$. In particular, we have $u^\gamma(\xi_1 + \xi_2) = u^\gamma(\xi_1) \cdot u^\gamma(\xi_2)$ if $\gamma \neq 0$ and $u^\gamma(\xi_1 + \xi_2) \equiv \xi_1 + \xi_2$ for $\gamma = 0$.

Then the utility assigned to the (random) outcome ξ is given by

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

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

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

In what follows, 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 compact set $\mathcal{A}_i$ 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 selected, then state j is reached in the next transition with a given probability $p_{ij}(a)$ and one-stage transition cost c_{ij} will be accrued to such transition.

[1]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

A (Markovian) policy controlling the decision process is given by a sequence of decisions (actions) at every time point. In particular, policy controlling the process, $\pi = (f^0, f^1, \ldots)$, with $\pi^k = (f^k, f^{k+1}, \ldots)$ for $k = 1, 2, \ldots$, hence also $\pi = (f^0, f^1, \ldots, f^{k-1}, \pi^k)$ 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, hence X is a homogeneous Markov chain with transition probability matrix $P(f)$ whose ij-th element equals $p_{ij}(f_i)$.

Let $\xi_n^\alpha = \sum_{k=0}^{n-1} \alpha^k c_{X_k, X_{k+1}}$ with $\alpha \in (0,1)$, resp. $\xi_n = \sum_{k=0}^{n-1} c_{X_k, X_{k+1}}$, be the stream of α-discounted, resp. undiscounted, transition costs received in the n next transitions of the considered Markov chain X. Similarly let $\xi^{(m,n)}$ be reserved for the total (random) cost obtained from the mth up to the nth transition (obviously, $\xi_n = c_{X_0, X_1} + \xi^{(1,n)}$). Moreover, if the risk sensitivity $\gamma \neq 0$ then $\bar{u}^\gamma(\xi_n^\alpha) = (\text{sign } \gamma)\, u^\gamma(\xi_n^\alpha)$, resp. $\bar{u}^\gamma(\xi_n) = (\text{sign } \gamma)\, u^\gamma(\xi_n)$, is the (random) utility assigned to ξ_n^α, resp. to ξ_n. Observe that $\xi^\alpha := \lim_{n\to\infty} \xi_n^\alpha$ is well defined, hence $u^\gamma(\xi^\alpha) = \exp(\gamma \sum_{k=0}^{\infty} \alpha^k c_{X_k, X_{k+1}})$.

In the overwhelming literature on stochastic dynamic programming attention was mostly paid to the risk neutral case, i.e. if $\gamma = 0$. The following results and techniques adapted from [8] and [12] will be useful for derivation of necessary and sufficient average optimality conditions and for their further extensions to risk-sensitive models.

Introducing for arbitrary $g, w_j \in \mathbb{R}$ $(i, j \in \mathcal{I})$ the discrepancy function

$$\tilde{\varphi}_{i,j}(w, g) = c_{i,j} - w_i + w_j - g \tag{3}$$

we can easily verify the following identities:

$$\xi_n^\alpha = \frac{1-\alpha^n}{1-\alpha} g + w_{X_0} - \alpha^n w_{X_n} + \sum_{k=0}^{n-1} \alpha^k [\tilde{\varphi}_{X_k, X_{k+1}}(w, g) - (1-\alpha) w_{X_{k+1}}] \tag{4}$$

$$\xi_n = ng + w_{X_0} - w_{X_n} + \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w, g). \tag{5}$$

If the process starts in state i and policy $\pi = (f^n)$ is followed then for the expected α-discounted or undiscounted total cost $V_i^\pi(\alpha, n) := \mathsf{E}_i^\pi\, \xi_n^\alpha$, $V_i^\pi(n) := \mathsf{E}_i^\pi\, \xi_n$ we immediately get by (4)–(5)

$$V_i^\pi(\alpha, n) = \frac{1-\alpha^n}{1-\alpha} g + w_i + \mathsf{E}_i^\pi \{\sum_{k=0}^{n-1} \alpha^k [\tilde{\varphi}_{X_k, X_{k+1}}(w, g) - (1-\alpha) w_{X_{k+1}}] - \alpha^n w_{X_n}\} \tag{6}$$

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

Observe that

$$\mathsf{E}_i^\pi \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w, g) = \sum_{j \in \mathcal{I}} p_{ij}(f_i) \{\tilde{\varphi}_{i,j}(w, g) + \mathsf{E}_j^{\pi^1} \sum_{k=1}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w, g)\}. \tag{8}$$

It is well-known from the dynamic programming literature (cf. e.g. [1, 6, 9, 10, 16]) that

If there exists state $i_0 \in \mathcal{I}$ that is accessible from any state $i \in \mathcal{I}$ for every $f \in \mathcal{F}$ then $(*)$

(i) For every $f \in \mathcal{F}$ the resulting transition probability matrix $P(f)$ is *unichain* (i.e. $P(f)$ have no two disjoint closed sets),

(ii) There exists decision vector $\hat{f} \in \mathcal{F}$ along with numbers $\hat{w}_i, i \in \mathcal{I}$ (unique up to additive constant), and $\hat{g}$ being the solution of the set of (nonlinear) equations

$$\hat{w}_i + \hat{g} = \min_{a \in \mathcal{A}_i} \sum_{j \in \mathcal{I}} p_{ij}(a)[c_{i,j} + \hat{w}_j] = \sum_{j \in \mathcal{I}} p_{ij}(\hat{f}_i)[c_{i,j} + \hat{w}_j], \tag{9}$$

$$\varphi_i(f, \hat{f}) := \sum_{j \in \mathcal{I}} p_{ij}(f)[c_{i,j} + \hat{w}_j] - \hat{w}_i - \hat{g} \geq 0 \quad \text{with} \quad \varphi_i(\hat{f}, \hat{f}) = 0. \tag{10}$$

From (6), (7), (10) we immediately get for $V_i^\pi(\alpha) := \lim_{n\to\infty} V_i^\pi(\alpha, n)$

$$V_i^{\hat\pi}(\alpha) = \frac{1}{1-\alpha}\hat g + \hat w_i - (1-\alpha)\mathsf{E}_i^{\hat\pi}\sum_{k=0}^{\infty}\alpha^k \hat w_{X_{k+1}}, \quad V_i^{\hat\pi}(n) = n\hat g + \hat w_i - \mathsf{E}_i^{\hat\pi}\hat w_n$$

hence for stationary policy $\pi \sim (\hat f)$ and arbitrary policy $\pi = (f^n)$

$$\lim_{n\to\infty}\frac{1}{n}V_i^{\hat\pi}(n) = \lim_{\alpha\uparrow 1}(1-\alpha)V_i^{\hat\pi}(\alpha) = \hat g \tag{11}$$

$$\lim_{n\to\infty}\frac{1}{n}V_i^{\pi}(n) = \hat g = \lim_{\alpha\uparrow 1}(1-\alpha)V_i^{\pi}(\alpha) \quad \text{if and only if}$$

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{E}_i^{\pi}\sum_{k=0}^{n-1}\varphi_{X_k}(f^n,\hat f) = 0. \tag{12}$$

2 Risk-Sensitive Optimality

On inserting the discrepancy function given by (3) in the exponential function $u^\gamma(\cdot)$ by (4) we get for the stream of discounted costs

$$\begin{aligned} u^\gamma(\xi_n^\alpha) &= \mathrm{e}^{\gamma\sum_{k=0}^{n-1}\alpha^k c_{X_k,X_{k+1}}} \\ &= \mathrm{e}^{\gamma[\sum_{k=0}^{n-1}\alpha^k g + w_{X_0} - \alpha^n w_{X_n}]} \times \mathrm{e}^{\gamma\sum_{k=0}^{n-1}\alpha^k[\tilde\varphi_{X_k,X_{k+1}}(w,g) - (1-\alpha)w_{X_{k+1}}]} \end{aligned} \tag{13}$$

and for $U_i^\pi(\gamma,\alpha,n) := \mathsf{E}_i^\pi u^\gamma(\xi_n^\alpha)$ we have

$$U_i^\pi(\gamma,\alpha,n) = \mathrm{e}^{\gamma[\frac{1-\alpha^n}{1-\alpha}g + w_i]} \times \mathsf{E}_i^\pi \mathrm{e}^{\gamma\{\sum_{k=0}^{n-1}\alpha^k[\tilde\varphi_{X_k,X_{k+1}}(w,g) - (1-\alpha)w_{X_{k+1}}] - \alpha^n w_{X_n}\}} \tag{14}$$

Observe that w_i's are bounded, i.e. $|w_{X_k}| \le K$ for some $K \ge 0$. Hence it holds

$$\mathrm{e}^{-|\gamma|K} \le \mathrm{e}^{\gamma(1-\alpha)\sum_{k=1}^{\infty}\alpha^k w_{X_{k+1}}} \le \mathrm{e}^{|\gamma|K} \tag{15}$$

and for n tending to infinity from (14) we immediately get for $U_i^\pi(\gamma,\alpha) := \lim_{n\to\infty} U_i^\pi(\gamma,\alpha,n)$

$$U_i^\pi(\gamma,\alpha) = \mathrm{e}^{\gamma[\frac{1}{1-\alpha}g + w_i]} \times \mathsf{E}_i^\pi \mathrm{e}^{\gamma\sum_{k=0}^{\infty}\alpha^k[\tilde\varphi_{X_k,X_{k+1}}(w,g) + (1-\alpha)w_{X_{k+1}}]} \tag{16}$$

Similarly for undiscounted models we get by (13), (14)

$$U_i^\pi(\gamma,n) = \mathrm{e}^{\gamma[ng + w_i]} \times \mathsf{E}_i^\pi \mathrm{e}^{\gamma[\sum_{k=0}^{n-1}\tilde\varphi_{X_k,X_{k+1}}(w,g) - w_{X_n}]} \tag{17}$$

Now observe that

$$\mathsf{E}_i^\pi \mathrm{e}^{\gamma\sum_{k=0}^{n-1}\tilde\varphi_{X_k,X_{k+1}}(w,g)} = \sum_{j\in\mathcal{I}} p_{ij}(f_i^0)\,\mathrm{e}^{\gamma[c_{i,j} - w_i + w_j - g]} \times \mathsf{E}_j^{\pi^1} \mathrm{e}^{\gamma\sum_{k=1}^{n-1}\tilde\varphi_{X_k,X_{k+1}}(w,g)} \tag{18}$$

$$\mathsf{E}_j^\pi\{\mathrm{e}^{\gamma\sum_{k=m}^{n-1}\tilde\varphi_{X_k,X_{k+1}}(w,g)} | X_m = j\} = \sum_{\ell\in\mathcal{I}} p_{j\ell}(f_j^m)\,\mathrm{e}^{\gamma[c_{j,\ell} - w_j + w_\ell - g]} \times \mathsf{E}_\ell^{\pi^{m+1}} \mathrm{e}^{\gamma\sum_{k=m+1}^{n-1}\tilde\varphi_{X_k,X_{k+1}}(w,g)} \tag{19}$$

Employing (18) the following facts can be easily verified by (16), (17).

Result 1.

(i) Let for a given stationary policy $\pi \sim (f)$ there exist $g(f)$, $w_j(f)$'s such that

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)\ \mathrm{e}^{\gamma[c_{i,j}-w_i(f)+w_j(f)-g(f)]} \ = \ \sum_{j\in\mathcal{I}} p_{ij}(f_i)\ \mathrm{e}^{\gamma\,\tilde{\varphi}_{i,j}(w(f),g(f))} = 1, \quad i\in\mathcal{I}. \tag{20}$$

Then

$$U_i^{\pi}(\gamma,\alpha) \ = \ \mathrm{e}^{\gamma(\frac{1}{1-\alpha}g(f)+w_i(f))} \times \mathsf{E}_i^{\,\pi}\,\mathrm{e}^{-\gamma\,(1-\alpha)\,\sum\limits_{k=1}^{\infty}\alpha^k\,w_{X_k}(f)} \tag{21}$$

$$U_i^{\pi}(\gamma,n) \ = \ \mathrm{e}^{\gamma(ng(f)+w_i(f))} \times \mathsf{E}_i^{\,\pi}\,\mathrm{e}^{-\gamma\ w_{X_n}(f)} \tag{22}$$

(ii) If it is possible to select $g=g^*$ and $w_j=w_j^*$'s, resp. $g=\hat{g}$ and $w_j=\hat{w}_j$'s, such that for any $f\in\mathcal{F}$, all $i\in\mathcal{I}$ and some $f^*\in\mathcal{F}$, resp. $\hat{f}\in\mathcal{F}$,

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)\ \mathrm{e}^{\gamma\,\tilde{\varphi}_{i,j}(w^*,g^*)} \le 1 \ \text{ with } \ \sum_{j\in\mathcal{I}} p_{ij}(f_i^*)\ \mathrm{e}^{\gamma\,\tilde{\varphi}_{i,j}(w^*,g^*)} = 1 \tag{23}$$

respectively

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)\ \mathrm{e}^{\gamma\,\tilde{\varphi}_{i,j}(\hat{w},\hat{g})} \ge 1 \ \text{ with } \ \sum_{j\in\mathcal{I}} p_{ij}(\hat{f}_i)\ \mathrm{e}^{\gamma\,\tilde{\varphi}_{i,j}(\hat{w},\hat{g})} = 1 \tag{24}$$

then

$$\mathrm{e}^{\gamma(\frac{1}{1-\alpha}\hat{g}+\hat{w}_i)}\cdot\mathrm{e}^{-|\gamma|K} \le U_i^{\hat{\pi}}(\gamma,\alpha) \le \ U_i^{\pi^*}(\gamma,\alpha) \le \ \mathrm{e}^{\gamma(\frac{1}{1-\alpha}g^*+w_i^*)}\cdot\mathrm{e}^{|\gamma|K} \tag{25}$$

$$\mathrm{e}^{\gamma(n\hat{g}+\hat{w}_i)}\cdot\mathrm{e}^{-|\gamma|K} \le U_i^{\hat{\pi}}(\gamma,n) \le \ U_i^{\pi^*}(\gamma,n)) \le \ \mathrm{e}^{\gamma(ng^*+w_i^*)}\cdot\mathrm{e}^{|\gamma|K}. \tag{26}$$

Result 2. Let (cf. (2)) $\quad Z_i^{\pi}(\gamma,\alpha)=\frac{1}{\gamma}\ln U_i^{\pi}(\gamma,\alpha), \quad Z_i^{\pi}(\gamma,n)=\frac{1}{\gamma}\ln U_i^{\pi}(\gamma,n).$

Then by (25), (26) for stationary policies $\hat{\pi}\sim(\hat{f})$, $\pi^*\sim(f^*)$, and by (16), (17) for an arbitrary policy $\pi=(f^n)$

$$\lim_{n\to\infty}\frac{1}{n}Z_i^{\hat{\pi}}(\gamma,n)=\lim_{\alpha\uparrow 1}(1-\alpha)Z_i^{\hat{\pi}}(\gamma,\alpha)=\hat{g}, \qquad \lim_{n\to\infty}\frac{1}{n}Z_i^{\pi^*}(\gamma,n)=\lim_{\alpha\uparrow 1}(1-\alpha)Z_i^{\pi^*}(\gamma,\alpha)=g^* \tag{27}$$

$$\lim_{n\to\infty}\frac{1}{n}Z_i^{\pi}(\gamma,n)=g^*, \quad \text{resp.} \quad \lim_{n\to\infty}\frac{1}{n}Z_i^{\pi}(\gamma,n)=\hat{g}, \quad \text{if and only if}$$

$$\lim_{n\to\infty}\frac{1}{n}\ln[\mathsf{E}_i^{\,\pi}\,\mathrm{e}^{\gamma\sum\limits_{k=0}^{n-1}\tilde{\varphi}_{X_k,X_{k+1}}(w^*,g^*)}]=0, \quad \text{resp.} \quad \lim_{n\to\infty}\frac{1}{n}\ln[\mathsf{E}_i^{\,\pi}\,\mathrm{e}^{\gamma\sum\limits_{k=0}^{n-1}\tilde{\varphi}_{X_k,X_{k+1}}(\hat{w},\hat{g})}]=0. \tag{28}$$

3 Poissonian Equations

The system of equations (20) for the considered stationary policy $\pi\sim(f)$ and the nonlinear systems of equations (23), (24) for finding stationary policy with maximal/minimal value of $g(f)$ can be also written as

$$\mathrm{e}^{\gamma[g(f)+w_i(f)]} \ = \ \sum_{j\in\mathcal{I}} p_{ij}(f_i)\,\mathrm{e}^{\gamma[c_{i,j}+w_j(f)]} \quad (i\in\mathcal{I}) \tag{29}$$

$$\mathrm{e}^{\gamma[g^*+w_i^*]} \ = \ \max_{f\in\mathcal{F}}\sum_{j\in\mathcal{I}} p_{ij}(f_i)\ \mathrm{e}^{\gamma[c_{i,j}+w_j^*]} \quad (i\in\mathcal{I}) \tag{30}$$

$$\mathrm{e}^{\gamma[\hat{g}+\hat{w}_i]} \ = \ \min_{f\in\mathcal{F}}\sum_{j\in\mathcal{I}} p_{ij}(f_i)\,\mathrm{e}^{\gamma[c_{i,j}+\hat{w}_j]} \quad (i\in\mathcal{I}) \tag{31}$$

respectively, for the values $g(f)$, $\hat{g}$, g^*, $w_i(f)$, w_i^*, $\hat{w}_i$ $(i=1,\dots,N)$; obviously, these values depend on the selected risk sensitivity γ. Eqs. (30), (31) can be called the *γ-average reward/cost optimality equation*. In particular, if $\gamma\downarrow 0$ using the Taylor expansion by (29), resp. (31), we have

$$g(f)+w_i(f)=\sum_{j\in\mathcal{I}} p_{ij}(f_i)\,[c_{i,j}+w_j(f)], \quad \text{resp.} \quad \hat{g}+\hat{w}_i=\min_{f\in\mathcal{F}}\sum_{j\in\mathcal{I}} p_{ij}(f_i)\,[c_{i,j}+\hat{w}_j]$$

that well corresponds to (9).

On introducing new variables $v_i(f) := \mathrm{e}^{\gamma w_i(f)}$, $\rho(f) := \mathrm{e}^{\gamma g(f)}$, and on replacing transition probabilities $p_{ij}(f_i)$'s by general nonnegative numbers defined by $q_{ij}(f_i) := p_{ij}(f_i) \cdot \mathrm{e}^{\gamma c_{ij}}$ (29) can be alternatively written as the following set of equations

$$\rho(f)v_i(f) \quad = \quad \sum_{j\in\mathcal{I}} q_{ij}(f_i)\, v_j(f) \quad (i \in \mathcal{I}) \tag{32}$$

and (30), (31) can be rewritten as the following sets of nonlinear equations (here $\hat{v}_i := \mathrm{e}^{\gamma \hat{w}_i}$, $v_i^* := \mathrm{e}^{\gamma w_i^*}$, $\hat{\rho} = \mathrm{e}^{\gamma \hat{g}}$, $\rho^* := \mathrm{e}^{\gamma g^*}$)

$$\rho^* v_i^* = \max_{f\in\mathcal{F}} \sum_{j\in\mathcal{I}} q_{ij}(f_i)\, v_j^*, \quad \hat{\rho}\,\hat{v}_i = \min_{f\in\mathcal{F}} \sum_{j\in\mathcal{I}} q_{ij}(f_i)\,\hat{v}_j \quad (i \in \mathcal{I}) \tag{33}$$

called *γ-average reward/cost optimality equation in multiplicative form.*

For what follows it is convenient to consider (32), (33) in matrix form. To this end, we introduce the $N \times N$ matrix $Q(f) = [q_{ij}(f_i)]$ with spectral radius (Perron eigenvalue) $\rho(f)$ along with its right Perron eigenvector $v(f) = [v_i(f_i)]$, hence (cf. [5]) $\rho(f)v(f) = Q(f)v(f)$. Similarly, for $v(f^*) = v^*$, $v(\hat{f}) = \hat{v}$ (33) can be written in matrix form as

$$\rho^* v^* = \max_{f\in\mathcal{F}} Q(f)v^*, \qquad \hat{\rho}\,\hat{v} = \min_{f\in\mathcal{F}} Q(f)\hat{v}. \tag{34}$$

Recall that vectorial maximum and minimum in (34) should be considered componentwise and $\hat{v}$, v^* are unique up to multiplicative constant.

Furthermore, if the transition probability matrix $P(f)$ is irreducible then also $Q(f)$ is irreducible and the right Perron eigenvector $v(f)$ can be selected strictly positive. To extend this assertion to unichain models in contrast to condition $(*)$ for the risk neutral case it is necessary to assume existence of state

$i_0 \in \mathcal{I}$ accessible from any state $i \in \mathcal{I}$ for every $f \in \mathcal{F}$ that belongs to the *basic* class[1] of $Q(f)$. $(**)$

If condition $(**)$ is fulfilled it can be shown (cf. [15]) that in (30), (31) and (33) eigenvectors $v(f)$, $\hat{v}$, v^* can be selected strictly positive and ρ^*, resp. $\hat{\rho}$, is the maximum, resp. minimum, Perron eigenvalue of the matrix family $\{Q(f), f \in \mathcal{F}\}$.

So we have arrived to

Result 3. Sufficient condition for the existence of γ-average reward/costs optimality equation is the existence of state $i_0 \in \mathcal{I}$ fulfilling condition $(**)$ (trivially fulfilled for irreducible models).

In particular, for unichain models condition $(**)$ is fulfilled if this risk sensitive coefficient γ is sufficiently close to zero (cf. [3, 4, 15]). Finding solution of (34) can be performed by policy or value iteration. Details can be found e.g. in [2, 3, 7, 14, 15].

Finally, we rewrite optimality condition (28) of Result 2 in terms of $Q(f)$, $\hat{v}$, $\hat{\rho}$. To this end first observe that

$$\begin{aligned} \mathsf{E}_{i_0}^{\pi}\, \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(\hat{w},\hat{g})} &= \prod_{k=0}^{n-1} \left\{ \sum_{i_{k+1}\in\mathcal{I}} p_{i_k,i_{k+1}}(f_{i_k}^k)\, \mathrm{e}^{\gamma[c_{i_k,i_{k+1}} + \hat{w}_{i_{k+1}} - \hat{w}_{i_k} - \hat{g}]} \right\} \\ &= \prod_{k=0}^{n-1} \left\{ \sum_{i_{k+1}\in\mathcal{I}} q_{i_k,i_{k+1}}(f_{i_k}^k) \cdot \hat{v}_{i_{k+1}} \cdot \hat{v}_{i_k}^{-1} \cdot \hat{\rho}^{-1} \right\} \end{aligned} \tag{35}$$

So equation (28) can be also written in matrix form as

$$\lim_{n\to\infty} \frac{1}{n} \ln \left\{ \prod_{k=0}^{n-1} \hat{V}^{-1} Q(f^k) \hat{V} \hat{\rho}^{-1} \right\} = \lim_{n\to\infty} \frac{1}{n} \ln \left\{ \hat{V}^{-1} \left[\prod_{k=0}^{n-1} Q(f^k) \cdot \hat{\rho}^{-1} \right] \cdot \hat{V} \right\} \cdot \hat{V} = \mathbf{0} \tag{36}$$

where the diagonal matrix $\hat{V} = \mathrm{diag}\,[\hat{v}_i]$ and $\mathbf{0}$ is reserved for a null matrix.

[1] (i.e. irreducible class with spectral radius equal to the Perron eigenvalue of $Q(f)$)

4 Conclusions

In this note necessary and sufficient optimality conditions for discrete time Markov decision chains are obtained along with equations for average optimal policies both for risk-neutral and risk-sensitive models. Our analysis is restricted to unichain models, and for the risk-sensitive case some additional assumptions are made. For multichain models it is necessary to find suitable partition of the state space into nested classes that retain some properties of the unichain model. Some results in this direction can be found in [11, 15, 17, 18].

Acknowledgements

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

References

[1] Bertsekas, D. P.: *Dynamic Programming and Optimal Control. Volume 2. Third Edition.* Athena Scientific, Belmont, Mass. 2007.

[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] 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.

[4] Cavazos-Cadena, R. and Hernández-Hernández, D.: A characterization of the optimal risk-sensitive average cost infinite controlled Markov chains. *Annals of Applied Probability* **15** (2005), 175–212.

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

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

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

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

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

[10] Ross, S. M.: *Introduction to Stochastic Dynamic Programming.* Academic Press, New York 1983.

[11] Rothblum, U. G. and Whittle, P.: Growth optimality for branching Markov decision chains. *Mathematics of Operations Research* **7** (1982), 582–601.

[12] Sladký, K.: On the set of optimal controls for Markov chains with rewards. *Kybernetika* **10** (1974), 526–547.

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

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

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

[16] Tijms, H. C.: *A First Course in Stochastic Models.* Wiley, Chichester, 2003.

[17] Whittle, P.: *Optimization Over Time – Dynamic Programming and Stochastic Control. Volume II, Chapter 35.* Wiley, Chichester 1983.

[18] Zijm, W. H. M.: *Nonnegative Matrices in Dynamic Programming.* Mathematical Centre Tract, Amsterdam 1983.