Cumulative Optimality in Risk-Sensitive and Risk-Neutral Markov Reward Chains

Karel Sladký[1]

Abstract.

This contribution is devoted to risk-sensitive and risk-neutral optimality in Markov decision chains. Since the traditional optimality criteria (e.g. discounted or average rewards) cannot reflect the variability-risk features of the problem, and using the mean variance selection rules that stem from the classical work of Markowitz present some technical difficulties, we are interested in expectation of the stream of rewards generated by the Markov chain that is evaluated by an exponential utility function with a given risk sensitivity coefficient. Recall that for the risk sensitivity coefficient equal zero we arrive at traditional optimality criteria. In this note we present necessary and sufficient risk-sensitivity and risk-neutral optimality conditions; in detail for unichain models and indicate their generalizations to multichain Markov reward chains.

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

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

1 Notation and Preliminaries

The usual optimization criteria examined in the literature on stochastic dynamic programming, such as a total discounted or mean (average) reward structures, may be quite insufficient to characterize the problem from the point of a decision maker. To this end it may be preferable if not necessary to select more sophisticated criteria that also reflect the variability-risk features of the problem. Perhaps the best known approaches stem from the classical work of Markowitz on mean variance selection rules. On the other hand risky decisions can be also eliminated when expectation of the stream of one stage rewards (or costs) is evaluated by an exponential utility function. Recall that exponential utility functions are separable and hence suitable for sequential decisions.

In what follows, we consider Markov decision chain $X = \{X_n,\, n = 0, 1, \ldots\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$ and an infinite (compact) set $\mathcal{A}_i \equiv [0, K_i] \subset \mathbb{R}$ 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 reward $r_{ij}(a) > 0$ will be accrued to such transition. We assume that each $p_{ij}(a)$, $r_{ij}(a)$ is a continuous function of $a \in \mathcal{A}_i$.

A (Markovian) policy controlling the chain, $\pi = (f^0, f^1, \ldots)$, 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 $f^n_i \in \mathcal{A}_i$ is the decision (or action) taken at the nth transition if the chain X is in state i. Let π^k be a sequence of decision vectors starting at the k-th transition, hence $\pi = (f^0, f^1, \ldots f^{k-1}, \pi^k)$. Policy which selects at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary; $P(f)$ is transition probability matrix with elements $p_{ij}(f_i)$. Stationary policy $\tilde{\pi}$ is randomized if there exist decision vectors $f^{(1)}, f^{(2)}, \ldots, f^{(m)} \in \mathcal{F}$ and on following policy $\tilde{\pi}$ we select in state i action $f^{(j)}_i$ with a given probability $\kappa^{(j)}_i$ (of course, $\kappa^{(j)}_i \geq 0$ with $\sum_{j=1}^{N} \kappa^{(j)}_i = 1$ for all $i \in \mathcal{I}$.)

[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

Let ξ_n be the cumulative reward obtained in the n first transition of the considered Markov chain X. Since the process starts in state X_0, $\xi_n = \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}$. Similarly let $\xi_{(m,n)}$ be reserved for the cumulative (random) reward, obtained from the mth up to the nth transition (obviously, $\xi_n = r_{X_0,X_1} + \xi_{(1,n)}$, we tacitly assume that $\xi_{(1,n)}$ starts in state X_1).

In this note, we assume that the stream of rewards 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 $\gamma \in \mathbb{R}$. Then the utility assigned to the (random) outcome ξ is given by

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

Obviously $\bar{u}^\gamma(\cdot)$ is continuous and strictly increasing. 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 $\gamma = 0$.

Moreover, recall that the certainty equivalent corresponding to ξ, say $Z^\gamma(\xi)$, is given by

$$\bar{u}^\gamma(Z^\gamma(\xi)) = \mathsf{E}\,[\bar{u}^\gamma(\xi)] \qquad \text{(the symbol } \mathsf{E} \text{ is reserved for expectation).} \tag{2}$$

From (1), (2) we can immediately conclude that

$$Z^\gamma(\xi) = \begin{cases} \gamma^{-1} \ln\{\mathsf{E}\; u^\gamma(\xi)\}, & \text{if } \gamma \neq 0 \\ \mathsf{E}\,[\xi] & \text{for } \gamma = 0. \end{cases} \tag{3}$$

Considering Markov decision process X, then if the process starts in state i, i.e. $X_0 = i$ and policy $\pi = (f^n)$ is followed, for the expectation of utility assigned to (cumulative) random reward ξ_n obtained in the n first transitions we get by (1)

$$\mathsf{E}_i^\pi\, \bar{u}^\gamma(\xi_n) := \begin{cases} (\text{sign } \gamma)\, \mathsf{E}_i^\pi\, \exp(\gamma\xi_n), & \text{if } \gamma \neq 0, \quad \text{risk-sensitive case} \\ \mathsf{E}_i^\pi\, \xi_n \quad \text{for } \gamma = 0 & \text{risk-neutral case.} \end{cases} \tag{4}$$

In what follows let

$$\bar{U}_i^\pi(\gamma, n) := \mathsf{E}_i^\pi\, \bar{u}^\gamma(\xi_n), \qquad U_i^\pi(\gamma, n) := \mathsf{E}_i^\pi\, \exp(\gamma\xi_n), \qquad V_i^\pi(n) := \mathsf{E}_i^\pi\, \xi_n. \tag{5}$$

2 Risk-Neutral Optimality in Markov Processes

In this section we focus attention primarily on so called unichain models, i.e. when the underlying Markov chain contains a single class of recurrent states. Then on introducing for arbitrary $g, w_j \in \mathbb{R}$ $(i, j \in \mathcal{I})$ the discrepancy function (cf. [8])

$$\tilde{\varphi}_{i,j}(w, g) := r_{ij} - w_i + w_j - g \tag{6}$$

we can easily verify the following identity:

$$\xi_n \;=\; ng + w_{X_0} - w_{X_n} + \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k,X_{k+1}}(w, g). \tag{7}$$

For the risk-neutral models (i.e. if the risk sensitivity coefficient $\gamma = 0$, and $u^\gamma(\xi) = \xi$) we can conclude:

If the process starts in state i and policy $\pi = (f^n)$ is followed then for the expected (undiscounted) total reward $V_i^\pi(n) := \mathsf{E}_i^\pi\, \xi_n$ we immediately get by (7)

$$V_i^\pi(n) \;=\; ng + w_i + \mathsf{E}_i^\pi\{\sum_{k=0}^{n-1} \tilde{\varphi}_{X_k,X_{k+1}}(w, g) - w_{X_n}\}, \qquad \text{where} \tag{8}$$

$$\mathsf{E}_i^\pi \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k,X_{k+1}}(w, g) \;=\; \sum_{j\in\mathcal{I}} p_{ij}(f_i)\{\tilde{\varphi}_{i,j}(w, g) + \mathsf{E}_j^{\pi^1} \sum_{k=1}^{n-1} \tilde{\varphi}_{X_k,X_{k+1}}(w, g)\} \tag{9}$$

It is well-known from the dynamic programming literature (cf. e.g. [1, 6, 9, 10]) 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. $P(f)$ has no two disjoint closed sets),

(ii) For every $f \in \mathcal{F}$ there exist numbers $g(f)$, and $w_i(f), i \in \mathcal{I}$ (unique up to additive constant) such that

$$w_i(f) + g(f) = \sum_{j \in \mathcal{I}} p_{ij}(f_i)[r_{ij} + w_j(f)], \qquad (i \in \mathcal{I}) \tag{10}$$

$$\text{i.e.} \quad \sum_{j \in \mathcal{I}} p_{ij}(f_i)\, \tilde{\varphi}_{i,j}(w,g) = 0 \quad \text{if} \quad \tilde{\varphi}_{i,j}(w,g) := r_{ij} - w_i(f) + w_j(f) - g(f).$$

(iii) There exists decision $\hat{f} \in \mathcal{F}$ (resp. $f^* \in \mathcal{F}$) along with numbers $\hat{g}$, (resp. g^*), $\hat{w}_i, i \in \mathcal{I}$ (resp. $w_i^*, i \in \mathcal{I}$) (unique up to additive constant) such that

$$\hat{w}_i + \hat{g} = \min_{a \in \mathcal{A}_i} \sum_{j \in \mathcal{I}} p_{ij}(a)[r_{ij} + \hat{w}_j] = \sum_{j \in \mathcal{I}} p_{ij}(\hat{f}_i)[r_{ij} + \hat{w}_j], \tag{11}$$

$$\varphi_i(f, \hat{f}) := \sum_{j \in \mathcal{I}} p_{ij}(f)[r_{ij} + \hat{w}_j] - \hat{w}_i - \hat{g} \geq 0 \ \text{ with } \ \varphi_i(\hat{f}, \hat{f}) = 0, \tag{12}$$

resp.

$$w_i^* + g^* = \max_{a \in \mathcal{A}_i} \sum_{j \in \mathcal{I}} p_{ij}(a)[r_{ij} + w_j^*] = \sum_{j \in \mathcal{I}} p_{ij}(f_i^*)[r_{ij} + w_j^*], \tag{13}$$

$$\varphi_i(f, f^*) := \sum_{j \in \mathcal{I}} p_{ij}(f)[r_{ij} + w_j^*] - w_i^* - g^* \leq 0 \ \text{ with } \ \varphi_i(f^*, f^*) = 0. \tag{14}$$

From (8),(10),(12),(14) we immediately get that $\hat{g} \leq g(f) \leq g^*$, and

$$V_i^{\hat{\pi}}(n) = n\hat{g} + \hat{w}_i - \mathsf{E}_i^{\hat{\pi}} \hat{w}_n, \qquad V_i^{\pi^*}(n) = ng^* + w_i^* - \mathsf{E}_i^{\pi^*} w_n^*. \tag{15}$$

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

$$\lim_{n\to\infty} \frac{1}{n} V_i^{\hat{\pi}}(n) = \lim_{n\to\infty} \frac{1}{n} V_i^{\pi}(n) = \hat{g} \quad \text{if and only if} \quad \lim_{n\to\infty} \frac{1}{n} \mathsf{E}_i^{\pi} \sum_{k=0}^{n-1} \varphi_{X_k}(f^n, \hat{f}) = 0. \tag{16}$$

$$\lim_{n\to\infty} \frac{1}{n} V_i^{\pi^*}(n) = \lim_{n\to\infty} \frac{1}{n} V_i^{\pi}(n) = g^* \quad \text{if and only if} \quad \lim_{n\to\infty} \frac{1}{n} \mathsf{E}_i^{\pi} \sum_{k=0}^{n-1} \varphi_{X_k}(f^n, f^*) = 0. \tag{17}$$

Remark. For the general multichain models it is necessary to modify $\tilde{\varphi}_{i,j}(w,g)$ such that $\tilde{\varphi}_{i,j}(w,g) := r_{ij} - w_i + w_j - g_i$ and introduce $\tilde{\psi}_{i,j}(g) := g_j - g_i$. Then (7) is replaced by

$$\xi_n = ng_{X_0} + w_{X_0} - w_{X_n} + \sum_{k=0}^{n-1} \Big[(n-1-k)\tilde{\psi}_{X_k, X_{k+1}}(g) + \tilde{\varphi}_{X_k, X_{k+1}}(w,g) \Big] \tag{18}$$

and (8) reads (see [12], [13])

$$V_i^{\pi}(n) = ng_i + w_i + \mathsf{E}_i^{\pi} \{ \sum_{k=0}^{n-1} \Big[(n-1-k)\, \tilde{\psi}_{X_k, X_{k+1}}(g) + \tilde{\varphi}_{X_k, X_{k+1}}(w,g) \Big] - w_{X_n} \}. \tag{19}$$

Then (10) should be completed with $\sum_{j \in \mathcal{I}} p_{ij}(f_i)[g_j(f) - g_i(f)] = 0$ and considered in the form $w_i(f) + g_i(f) = \sum_{j \in \mathcal{I}} p_{ij}(f_i)[r_{ij} + w_j(f)]$. Similarly optimal policy $\hat{\pi} \sim (\hat{f})$, $\pi^* \sim (f^*)$ must fulfil

$$\min_{a \in \mathcal{A}_i} \sum_{j \in \mathcal{I}} p_{ij}(a)[\hat{g}_j - \hat{g}_i] = \sum_{j \in \mathcal{I}} p_{ij}(\hat{f}_i)[\hat{g}_j - \hat{g}_i], \qquad \max_{a \in \mathcal{A}_i} \sum_{j \in \mathcal{I}} p_{ij}(a)[g_j^* - g_i^*] = \sum_{j \in \mathcal{I}} p_{ij}(f_i^*)[g_j^* - g_i^*] \tag{20}$$

and in (11) (resp. in (13)) minimization (resp. maximization) of $a \in \mathcal{A}_i$ should be considered only for $a \in \bar{\mathcal{A}}_i \subset \mathcal{A}_i$ fulfilling (20). Then, if *the action set is finite* it is guaranteed that for sufficiently large n $(n-1-k)\tilde{\psi}_{X_k, X_{k+1}}(\hat{g}) + \tilde{\varphi}_{X_k, X_{k+1}}(\hat{w}, \hat{g}) \leq 0$ (resp. $(n-1-k)\tilde{\psi}_{X_k, X_{k+1}}(g^*) + \tilde{\varphi}_{X_k, X_{k+1}}(w^*, g^*) \geq 0$).

3 Risk-Sensitive Optimality in Unichain Markov Processes

Similarly to risk-neutral models we get by (5), (6), (7) for the risk-sensitive case

$$U_i^{\pi}(\gamma, n) = \mathrm{e}^{\gamma[ng+w_i]} \times \mathsf{E}_i^{\pi} \mathrm{e}^{\gamma[\sum_{k=0}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w,g) - w_{X_n}]}. \tag{21}$$

Now observe that

$$\mathsf{E}_i^{\pi}\, \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w,g)} = \sum_{j \in \mathcal{I}} p_{ij}(f_i^0)\, \mathrm{e}^{\gamma[r_{ij} - w_i + w_j - g]} \times \mathsf{E}_j^{\pi^1}\, \mathrm{e}^{\gamma \sum_{k=1}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w,g)} \tag{22}$$

$$\mathsf{E}_j^{\pi} \{\mathrm{e}^{\gamma \sum_{k=m}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w,g)} | X_m = j\} = \sum_{\ell \in \mathcal{I}} p_{j,\ell}(f_j^m)\, \mathrm{e}^{\gamma[r_{j,\ell} - w_j + w_\ell - g]} \times \mathsf{E}_\ell^{\pi^{m+1}}\, \mathrm{e}^{\gamma \sum_{k=m+1}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w,g)}. \tag{23}$$

In analogy with the risk-neutral case if stationary policy $\pi \sim (f)$ is followed, we are looking for numbers g, w_j's such that $\sum_{j \in \mathcal{I}} p_{ij}(f_i)\, \mathrm{e}^{\gamma \tilde{\varphi}_{ij}(g,w)} = 1$ and for stationary policy with maximal/minimal value of $g(f)$. To this end we consider the following sets of linear and nonlinear equations

$$\mathrm{e}^{\gamma[g(f) + w_i(f)]} = \sum_{j \in \mathcal{I}} p_{ij}(f_i)\ \mathrm{e}^{\gamma[r_{ij} + w_j(f)]} \quad (i \in \mathcal{I}) \tag{24}$$

$$\mathrm{e}^{\gamma[g^* + w_i^*]} = \max_{f \in \mathcal{F}} \sum_{j \in \mathcal{I}} p_{ij}(f_i)\ \mathrm{e}^{\gamma[r_{ij} + w_j^*]}, \qquad \mathrm{e}^{\gamma[\hat{g} + \hat{w}_i]} = \min_{f \in \mathcal{F}} \sum_{j \in \mathcal{I}} p_{ij}(f_i)\ \mathrm{e}^{\gamma[r_{ij} + \hat{w}_j]} \quad (i \in \mathcal{I}) \tag{25}$$

for the values $g(f)$, $\hat{g}$, g^*, $w_i(f)$, w_i^*, $\hat{w}_i$ $(i = 1, \ldots, N)$; obviously, these values depend on the selected risk sensitivity γ. Eqs. (25) can be called the *γ-average reward/cost optimality equation*. In particular, if $\gamma \downarrow 0$ using the Taylor expansion by (24), resp. (25), we have

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

that well corresponds to (11).

On introducing the 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 $q_{ij}(f_i) := p_{ij}(f_i) \cdot \mathrm{e}^{\gamma r_{ij}}$ (24) can be alternatively written as the following set of equations

$$\rho(f) v_i(f) = \sum_{j \in \mathcal{I}} q_{ij}(f_i)\, v_j(f) \quad (i \in \mathcal{I}) \tag{26}$$

and (25) can be rewritten as the following sets of nonlinear equations (here $\hat{v}_i := \mathrm{e}^{\gamma \hat{w}_i}$, $v_i^* := \mathrm{e}^{\gamma w_i^*}$, $\hat{\rho} := \mathrm{e}^{\gamma \hat{g}}$, $\rho^* := \mathrm{e}^{\gamma g^*}$)

$$\rho^* v_i^* = \max_{f \in \mathcal{F}} \sum_{j \in \mathcal{I}} q_{ij}(f_i)\, v_j^*, \quad \hat{\rho}\, \hat{v}_i = \min_{f \in \mathcal{F}} \sum_{j \in \mathcal{I}} q_{ij}(f_i)\, \hat{v}_j \quad (i \in \mathcal{I}) \tag{27}$$

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

For what follows it is convenient to consider (26), (27) in matrix form. To this end we introduce (cf. [5]) $N \times N$ matrix $Q(f) = [q_{ij}(f_i)]$ with spectral radius (Perron eigenvalue) $\rho(f)$ along with its right Perron eigenvector $v(f) = [v_i(f)]$, and right Perron eigenvectors $v(f^*) = v^* = [v_i^*]$, $v(\hat{f}) = \hat{v} = [\hat{v}_i]$. Then (26),(27) can be written in matrix form as

$$\rho(f) v(f) = Q(f) v(f), \quad \rho^* v^* = \max_{f \in \mathcal{F}} Q(f) v^*, \quad \hat{\rho}\, \hat{v} = \min_{f \in \mathcal{F}} Q(f) \hat{v}. \tag{28}$$

Recall that vectorial maximum and minimum in (28) 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. Unfortunately, if $P(f)$ is unichain in contrast to condition $(*)$ to guarantee that $v(f)$ can be selected strictly positive 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], [16]) that

(i) In (28) 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 $\{Q(f), f \in \mathcal{F}\}$.

(ii) From (3), (21), (22), (24) we immediately get for stationary policy $\pi \sim (f)$ that

$$U_i^{\pi}(\gamma, n) = \mathrm{e}^{\gamma[ng(f)+w_i(f)]} \times \mathsf{E}_i^{\pi} \mathrm{e}^{\gamma w_{X_n}(f)}, \qquad Z_i^{\pi}(\gamma, n) = \frac{1}{\gamma} \ln U_i^{\pi}(\gamma, n).$$

Similarly, for the mean value of the certainty equivalent for stationary policies $\hat{\pi} \sim (\hat{f})$, $\pi^* \sim (f^*)$, and an arbitrary policy $\pi = (f^n)$ we get

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

$$\lim_{n\to\infty} \frac{1}{n} \ln[\mathsf{E}_i^{\pi} \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \bar{\varphi}_{x_k, x_{k+1}}(w^*, g^*)}] = 0, \quad \text{resp.} \quad \lim_{n\to\infty} \frac{1}{n} \ln[\mathsf{E}_i^{\pi} \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \bar{\varphi}_{x_k, x_{k+1}}(\hat{w}, \hat{g})}] = 0. \quad (29)$$

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 (28) can be performed by policy or value iteration. Details can be found e.g. in [2, 3, 7, 14, 15, 16].

4 Risk-Sensitive Optimality in Multichain Markov Processes

To begin with, recall that a (reducible) nonnegative matrix $Q(f), f \in \mathcal{F}$ has only nonnegative (not necessary positive) right Perron eigenvectors. Then (cf. [11, 14, 15, 17, 18])

i) On suitably permuting rows and corresponding columns each $Q(f), f \in \mathcal{F}$ can be written in block-triangular form such that its (possible reducible) diagonal blocks of $Q(f)$, say $Q_{ii}(f)$, are the biggest submatrices of $Q(f)$ with strictly positive right Perron eigenvectors, i.e. for $Q_{ii}(f)$, the ith diagonal block of $Q(f)$ and for the corresponding (strictly positive) right Perron eigenvector $\bar{v}_i(f)$ it holds

$$\rho_i(f)\, \bar{v}_i(f) = Q_{ii}(f)\, \bar{v}_i(f), \quad \text{where} \quad \bar{v}_i(f) > 0 \ \text{ and } \ \rho_{i-1}(f) \le \rho_i(f) \le \rho_{i+1}(f). \quad (30)$$

Considering diagonal blocks $Q_{i-1,i-1}(f)$ and $Q_{ii}(f)$ where $\rho_{i-1}(f) = \rho_i(f)$ accessibility of basic classes of $Q_{i-1,i-1}(f)$ and $Q_{ii}(f)$ is of great importance.

ii) Considering the set of nonnegative matrices $Q(f), f \in \mathcal{F}$ (i.e. is a family of nonnegative matrices fulfilling the "product property") it is possible to construct (using policy iteration algorithms) the matrix $Q(\hat{f})$ (resp. $Q(f^*)$ whose diagonal blocks are the biggest submatrices with positive right Perron eigenvectors and minimum (resp. maximum) possible spectral radii of the set $Q(f), f \in \mathcal{F}$. In particular:

There exist $f^{\odot} = \hat{f}, f^* \in \mathcal{F}$ such that the matrix

$$Q(f^{\odot}) = \begin{bmatrix} Q_{11}(f^{\odot}) & Q_{12}(f^{\odot}) & \dots & Q_{1s}(f^{\odot}) \\ 0 & Q_{22}(f^{\odot}) & \dots & Q_{2s}(f^{\odot}) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Q_{ss}(f^{\odot}) \end{bmatrix} \quad (31)$$

induces the *basic partition* of the state space $\mathcal{I}$, such that for $f^{\odot} = \hat{f}, f^* \in \mathcal{F}$

$$\mathcal{I} = \mathcal{I}_1(f^{\odot}) \cup \mathcal{I}_2(f^{\odot}) \cup \ldots \cup \mathcal{I}_s(f^{\odot}), \quad \text{where} \quad \mathcal{I}_i(f^{\odot}) \cap \mathcal{I}_j(f^{\odot}) = \emptyset \quad \text{for} \quad i \neq j,$$

and in (31) elements of the $Q_{ii}(f^{\odot})$ are labelled from $\mathcal{I}_i(f^{\odot})$.

Furthermore, on keeping the basic partition given by $Q(f^*)$ then $Q_{ji}(f^*) \equiv 0$ for all $j < i$ and

$\rho_i(f^*)\, v_i(f^*) = Q_{ii}(f^*)\, \bar{v}_i(f^*) \ge Q_{ii}(f)\, \bar{v}_i(f^*)$, where $\bar{v}_i(f^*) > 0$ and $\rho_{i-1}(f^*) \le \rho_i(f^*) \le \rho_{i+1}(f^*)_i$ with $\rho_i(f^*) = \rho_{i+1}(f^*)$ if and only if each basic class of $Q_{ii}(f^*)$ has access to some basic class of $Q_{i+1,i+1}(f^*)$. On considering submatrices $Q_{ii}(f)$ with elements from $\mathcal{I}_i(f^{\odot})$ we can apply results of Section 3 to $Q_{ii}(f)$.

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

5 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 mostly restricted to unichain models, and for the risk-sensitive case some additional assumptions are made. If no such assumptions are made, it is indicated how to handle this problem by partition of the state space into suitable classes that inherit from properties of unichain models. Some further results in this direction can be found in [11, 14, 15, 17, 18].

Acknowledgements

This research was supported by the Czech Science Foundation under Grants P402/11/0150 and 13-14445S.

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, *Math. Oper. Res.* **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, *Math. Methods Oper. Res.* **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 Appl. Probab.* **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, *Manag. Sci.* **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, *Math. Oper. Res.* **7**(1982), 582–601.

[12] Sladký, K.: Necessary and sufficient optimality conditions for average reward of controlled Markov chains, *Kybernetika* **9** (1973), 124–137.

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

[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] Sladký, K.: Risk-sensitive and average optimality in Markov decision processes. In: *Proc. 30th Internat. Conference Mathematical Methods in Economics 2012, Part II* (J. Ramík and D. Stavárek, eds.), Silesian University, School of Business Administration, Karviná 2012, 799–804.

[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.