Kybernetika

Karel Sladký

# Risk-sensitive average optimality in Markov decision processes

*Kybernetika*, Vol. 54 (2018), No. 6, 1218–1230

Persistent URL: `http://dml.cz/dmlcz/147606`

KYBERNETIKA — VOLUME *54* (2018), NUMBER 6, PAGES 1218–1230

# RISK-SENSITIVE AVERAGE OPTIMALITY IN MARKOV DECISION PROCESSES

Karel Sladký

In this note attention is focused on finding policies optimizing risk-sensitive optimality criteria in Markov decision chains. To this end we assume that the total reward generated by the Markov process is evaluated by an exponential utility function with a given risk-sensitive coefficient. The ratio of the first two moments depends on the value of the risk-sensitive coefficient; if the risk-sensitive coefficient is equal to zero we speak on risk-neutral models. Observe that the first moment of the generated reward corresponds to the expectation of the total reward and the second central moment of the reward variance.

For communicating Markov processes and for some specific classes of unichain processes long run risk-sensitive average reward is independent of the starting state. In this note we present necessary and sufficient condition for existence of optimal policies independent of the starting state in unichain models and characterize the class of average risk-sensitive optimal policies.

*Keywords:* controlled Markov processes, finite state space, asymptotic behavior, risk-sensitive average optimality

*Classification:* 90C40, 93E20

## 1. INTRODUCTION

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 (cf. [19, 20]) on mean variance selection rules. Unfortunately, finding appropriate values of the minimal variance is a difficult problem in stochastic dynamic programming (cf. e. g. [17, 27]).

On the other hand risky decisions can be also eliminated in so-called risk sensitive models where 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 and observe that the variance can be easily approximated by expanding exponential utility function if the risk aversion coefficient is sufficiently small.

DOI: 10.14736/kyb-2018-6-1218

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 $\mathcal{A}_i \equiv [0, K_i] \subset \mathcal{R}$ of possible decisions (actions) in state $i \in \mathcal{I}$ if the $X$ is unichain[1]. 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 $r_{ij} > 0$ will be accrued to such transition. We assume that each $p_{ij}(a)$ 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 $f_i^n \in \mathcal{A}_i$ is the decision (or action) taken at the $n$th transition if the chain $X$ is in state $i$. Let $\pi^k$ 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 $f_i^{(j)}$ with a given probability $\kappa_i^{(j)}$ (of course, $\kappa_i^{(j)} \ge 0$ with $\sum_{j=1}^{N} \kappa_i^{(j)} = 1$ for all $i \in \mathcal{I}$).

Let $\xi_n$ be the cumulative reward obtained in the $n$ first transitions 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 $m$th up to the $n$th 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 $\gamma \in \mathcal{R}$. Then the utility assigned to the (random) outcome $\xi$ is given by

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

Obviously $\bar{u}^\gamma(\cdot)$ is continuous and strictly increasing. For $\gamma > 0$ (risk seeking case) $\bar{u}^\gamma(\cdot)$ is convex, if $\gamma < 0$ (risk averse 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 $\gamma \ne 0$ and additive for $\gamma = 0$. In particular, for $u^\gamma(\xi) := \exp(\gamma\xi)$ we have $u^\gamma(\xi_1 + \xi_2) = u^\gamma(\xi_1) \cdot u^\gamma(\xi_2)$ if $\gamma \ne 0$ and $u^\gamma(\xi_1 + \xi_2) \equiv \xi_1 + \xi_2$ for $\gamma = 0$.

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

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

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

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

[1] Notice that some problems can arise if no unichain assumption is made (cf. [2, 13]).

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 $\xi_n$ obtained in the $n$ first transitions we get by (1)

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

In what follows let

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

The paper is organized as follows. Section 2 summarized necessary and sufficient optimality condition for unichain risk-neutral Markov models with average optimality criteria where expectation of total reward is based on discrepancy functions. In Section 3 similar approach is developed for the risk sensitive optimality in Markov unichain models. Unfortunately in contrast to the risk-neutral unichain models if the underlying Markov chain is not communicating the value of the long run risk-average return need not be independent of the starting state. Section 4 presents necessary and sufficient condition for unichain models, as well as for a very specific multi-chain models, based on the Perron–Frobenius theory guaranteeing that the growth of expected utility is independent of the starting state. Conclusions are made in Section 5.

## 2. RISK-NEUTRAL OPTIMALITY IN MARKOV PROCESSES

In this section we focus attention on so called unichain models, i.e. when the underlying Markov chain contains a single class of recurrent states, and present characterization of control policies by discrepancy functions. Discrepancy functions were originally introduced in [17] for risk-neutral unichain models, unfortunately they were widely recognized later (see [18] or the review paper [1], page 319). To this end, on introducing for arbitrary $g, w_j \in \mathcal{R}$ $(i, j \in \mathcal{I})$ the discrepancy function

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

for the random reward obtained up to the $n$th transition we have

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

Hence by (5) for the expectation of $\xi_n$ (the risk-neutral case) we get

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

In what follows we make

**Assumption 1.** There exists state $i_0 \in \mathcal{I}$ that is accessible from any state $i \in \mathcal{I}$ for every $f \in \mathcal{F}$.

Obviously, if Assumption 1 holds, then the resulting transition probability matrix $P(f)$ is *unichain* for every $f \in \mathcal{F}$ (i. e. $P(f)$ has no two disjoint closed sets).

The following facts are well-known from the dynamic programming literature (cf. e. g. [15, 21, 22, 23, 24]).

(i) 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

$$w_i(f) + g(f) = \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} + w_j(f)], (i \in \mathcal{I}), \quad \text{i. e.} \tag{9}$$

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

(ii) 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

$$\hat{w}_i + \hat{g} \ = \ \min_{a\in\mathcal{A}_i} \sum_{j\in\mathcal{I}} p_{ij}(a)[r_{ij} + \hat{w}_j] = \sum_{j\in\mathcal{I}} p_{ij}(\hat{f}_i)[r_{ij} + \hat{w}_j], \tag{10}$$

$$\varphi_i(f,\hat{f}) \ := \ \sum_{j\in\mathcal{I}} p_{ij}(f)[r_{ij} + \hat{w}_j] - \hat{w}_i - \hat{g} \ge 0 \ \text{ with } \ \varphi_i(\hat{f},\hat{f}) = 0, \tag{11}$$

resp.

$$w^*_i + g^* \ = \ \max_{a\in\mathcal{A}_i} \sum_{j\in\mathcal{I}} p_{ij}(a)[r_{ij} + w^*_j] = \sum_{j\in\mathcal{I}} p_{ij}(f^*_i)[r_{ij} + w^*_j], \tag{12}$$

$$\varphi_i(f,f^*) \ := \ \sum_{j\in\mathcal{I}} p_{ij}(f)[r_{ij} + w^*_j] - w^*_i - g^* \le 0 \ \text{ with } \ \varphi_i(f^*,f^*) = 0. \tag{13}$$

From (8),(9),(11),(12) we immediately get that $\hat{g} \le g(f) \le g^*$, and for stationary policies $\hat{\pi} \sim (\hat{f}), \pi^* \sim (f^*)$

$$V^{\hat{\pi}}_i(n) = n\hat{g} + \hat{w}_i - \mathrm{E}^{\hat{\pi}}_i \hat{w}_n, \qquad V^{\pi^*}_i(n) = ng^* + w^*_i - \mathrm{E}^{\pi^*}_i w^*_n. \tag{14}$$

Moreover, for arbitrary policy $\pi = (f^n)$ it holds

$$\hat{g} \le \liminf_{n\to\infty} \tfrac{1}{n} V^{\pi}_i(n) \le \limsup_{n\to\infty} \tfrac{1}{n} V^{\pi}_i(n) \le g^* \ \text{, and}$$

$$\lim_{n\to\infty} \frac{1}{n} V^{\pi}_i(n) = \hat{g} \qquad \text{if and only if} \quad \lim_{n\to\infty} \frac{1}{n} \mathrm{E}^{\pi}_i \sum_{k=0}^{n-1} \varphi_{X_k}(f^n,\hat{f}) = 0, \tag{15}$$

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

Similarly,

$$\lim_{n\to\infty} |V^{\hat{\pi}}_i(n) - n\hat{g}| \qquad \text{resp.} \quad \lim_{n\to\infty} |V^{\pi^*}_i(n) - ng^*| \ \text{is bounded if and only if}$$

$$\mathrm{E}^{\pi}_i \sum_{k=0}^{n-1} \varphi_{X_k}(f^n,\hat{f}) \qquad \text{resp.} \quad \mathrm{E}^{\pi}_i \sum_{k=0}^{n-1} \varphi_{X_k}(f^n,f^*) \quad \text{is bounded.}$$

Moreover, from (8) we immediately also get

$$\lim_{n\to\infty}\frac{1}{n}\xi_n = g \text{ a.s. if and only if } \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\tilde{\varphi}_{X_k,X_{k+1}}(w,g) = 0 \text{ a.s.} \tag{17}$$

## 3. RISK-SENSITIVE OPTIMALITY IN UNICHAIN MARKOV PROCESSES

In this section we assume that the risk sensitivity coefficient $\gamma \neq 0$ and the transition probability matrix $P(f)$ is unichain for every $f \in \mathcal{F}$, i.e. Assumption 1 is fulfilled. Similarly to the risk-neutral models, let for real $g$, $w_i$'s $(i \in \mathcal{I})$

$$\tilde{\varphi}_{ij}(w,g) := r_{ij} - g + w_j - w_i, \quad \text{where } w' = \min_{i\in\mathcal{I}} w_i,\ w'' = \max_{i\in\mathcal{I}} w_i.$$

Then $r_{X_k,X_{k+1}} = \tilde{\varphi}_{X_k,X_{k+1}}(w,g) + g - w_{X_{k+1}} + w_{X_k}$ and if policy $\pi = (f^n)$ is followed we get by (5),(6),(7) for the risk-sensitive case

$$U_i^\pi(\gamma,n) = \mathrm{E}_i^\pi \mathrm{e}^{\gamma\sum\limits_{k=0}^{n-1} r_{X_k,X_{k+1}}} = \mathrm{e}^{\gamma[ng+w_i]} \times \mathrm{E}_i^\pi \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1}\tilde{\varphi}_{X_k,X_{k+1}}(w,g) - w_{X_n}]}. \tag{18}$$

The first term on the RHS of (18) is non-random and hence if $\gamma\, w'' > 0$

$$\mathrm{E}_i^\pi\, \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1}\tilde{\varphi}_{X_k,X_{k+1}}(w,g) - w'']} \le \frac{U_i^\pi(\gamma,n)}{\mathrm{e}^{\gamma[ng+w_i]}} \le \mathrm{E}_i^\pi\, \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1}\tilde{\varphi}_{X_k,X_{k+1}}(w,g) - w']} \tag{19}$$

and similarly for $\gamma\, w'' < 0$

$$\mathrm{E}_i^\pi\, \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1}\tilde{\varphi}_{X_k,X_{k+1}}(w,g) - w']} \le \frac{U_i^\pi(\gamma,n)}{\mathrm{e}^{\gamma[ng+w_i]}} \le \mathrm{E}_i^\pi\, \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1}\tilde{\varphi}_{X_k,X_{k+1}}(w,g) - w'']}. \tag{20}$$

Unfortunately, risk-sensitive analogies to (9),(10) and (12) as well as optimality conditions are more complicated and the unichain property itself (cf. Assumption 1) is not sufficient for the existence of $g, w_i$'s fulfilling (18)–(20).

In what follows we show that under certain conditions it is possible to select $w_i$'s and $g$ such that for stationary policy $\pi \sim (f)$ and any $i, j, X_k \in \mathcal{I}$

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)\, \mathrm{e}^{\gamma[r_{ij}+w_j]} = \mathrm{e}^{\gamma[g+w_i]} \quad \text{or} \quad \mathrm{E}_i^\pi\, \mathrm{e}^{\gamma\tilde{\varphi}_{X_k,X_{k+1}}(w,g)} = 1. \tag{21}$$

The first equation, called also the Poissonian equation, was discussed frequently in the literature and recognized as a sufficient condition for existence of the risk-sensitive optimality where the optimality condition does not depend on the starting state.

We shall start our analysis with the following

**Lemma 2.1.** Let (21) hold for stationary policy $\pi \sim (f)$. Then for $\tilde{w} = w', w''$

$$\mathrm{E}_i^\pi\, \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1}\tilde{\varphi}_{X_k,X_{k+1}}(w,g) - \tilde{w}]} = \mathrm{e}^{\gamma[-\tilde{w}]} \tag{22}$$

and hence

$$\mathrm{e}^{\gamma[ng+w_i-w'']} \le U_i^\pi(\gamma,n) \le \mathrm{e}^{\gamma[ng+w_i-w']} \quad \text{if} \quad \gamma w'' > 0, \tag{23}$$

$$\mathrm{e}^{\gamma[ng+w_i-w']} \le U_i^\pi(\gamma,n) \le \mathrm{e}^{\gamma[ng+w_i-w'']} \quad \text{if} \quad \gamma w'' < 0. \tag{24}$$

Proof. To verify (22) after some algebra and on employing (21) we conclude that

$$\begin{aligned}
&\mathrm{E}_i^\pi\, \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1} \tilde\varphi_{X_k,X_{k+1}}(w,g)-w']} = \mathrm{E}_i^\pi\, \mathrm{e}^{\gamma \sum\limits_{k=0}^{n-2} \tilde\varphi_{X_k,X_{k+1}}(w,g)+\gamma\tilde\varphi_{X_{n-1},X_n}(w,g)} \cdot \mathrm{e}^{-\gamma w'} \\
&= \sum_{\ell\in\mathcal{I}} p_{j,\ell}(f_j)\{\mathrm{E}_i^\pi\, [\mathrm{e}^{\gamma \sum\limits_{k=0}^{n-2} \tilde\varphi_{X_k,X_{k+1}}(w,g)} | X_{n-1}=j] \cdot \mathrm{e}^{\gamma[r_{j\ell}-w_j+w_\ell-g]}\} \cdot \mathrm{e}^{-\gamma w'} \\
&= \mathrm{E}_i^\pi\, \mathrm{e}^{\gamma \sum\limits_{k=0}^{n-2} \tilde\varphi_{X_k,X_{k+1}}(w,g)} \cdot \mathrm{e}^{-\gamma w'}
\end{aligned} \tag{25}$$

The proof for $w''$ goes on the same lines.

In particular, on iterating the above displayed formula we can conclude that (22) also holds for $\tilde w = w''$. Then on inserting (22) into (19),(20) we immediately get (23), (24). □

Observe that (18),(19),(20) are the risk-sensitive analogies of (8) for the risk-neutral case.[2] Since (cf.(23),(24)) for stationary policy $\pi \sim (f)$

$$\frac{1}{\gamma} \ln U_i^\pi(\gamma,n) = ng + w_i + h(n)$$

where $h(n)$ is bounded, in particular $|h(n)| \le \max\{|\gamma w'|, |\gamma w''|\}$.

In the next section attention is focused on finding necessary and sufficient condition for existence $g(f), w_i(f)$'s such that for stationary policy $\pi \sim (f)$ (21) holds. Furthermore, we are looking for stationary policy with maximal/minimal value of $g(f)$. As we shall see later, on comparing with the risk neutral case, the problem is more complicated since the unichain property of the underlying Markov process is not sufficient.

To this end, we shall consider the following sets of linear and nonlinear equations

$$\mathrm{e}^{\gamma[g(f)+w_i(f)]} = \sum_{j\in\mathcal{I}} p_{ij}(f_i)\, \mathrm{e}^{\gamma[r_{ij}+w_j(f)]} \quad (i\in\mathcal{I}) \tag{26}$$

$$\mathrm{e}^{\gamma[g^*+w_i^*]} = \max_{f\in\mathcal{F}} \sum_{j\in\mathcal{I}} p_{ij}(f_i)\, \mathrm{e}^{\gamma[r_{ij}+w_j^*]} \quad (i\in\mathcal{I}) \tag{27}$$

$$\mathrm{e}^{\gamma[\hat g+\hat w_i]} = \min_{f\in\mathcal{F}} \sum_{j\in\mathcal{I}} p_{ij}(f_i)\, \mathrm{e}^{\gamma[r_{ij}+\hat w_j]} \quad (i\in\mathcal{I}) \tag{28}$$

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 $\gamma$. Eqs. (27),(28) can be called the $\gamma$-*average reward/cost optimality equation.*

[2] In particular, from (8) we immediately get
$g + \frac{1}{n}\mathrm{E}_i^\pi\{\sum\limits_{k=0}^{n-1} \tilde\varphi_{X_k,X_{k+1}}(w,g)\} \le \frac{1}{n}V_i^\pi(n) \le g + \frac{1}{n}\mathrm{E}_i^\pi\{\sum\limits_{k=0}^{n-1} \tilde\varphi_{X_k,X_{k+1}}(w,g)\}.$

Necessary and sufficient conditions guaranteeing solutions of (26),(27),(28) are discussed in the next section. Now we are in a position to formulate necessary and sufficient average reward optimality conditions for the risk sensitive models if optimality equations (26),(27),(28) are fulfilled. Recall that the discrepancy function $\tilde{\varphi}_{x_k,x_{k+1}}(w,g) := r_{ij} - g + w_j - w_i$.

**Theorem 1.** Let $\hat{g} \le g^*$ be the solution of (27) and (28). Then for an arbitrary policy $\pi = (f^n)$

$$\hat{g} \le \liminf_{n\to\infty} \frac{1}{n} Z_i^\pi(\gamma,n) \le \limsup_{n\to\infty} \frac{1}{n} Z_i^\pi(\gamma,n) \le g^*, \quad \text{and}$$

$$\lim_{n\to\infty} \frac{1}{n} Z_i^\pi(\gamma,n) = g^* \Longleftrightarrow \lim_{n\to\infty} \frac{1}{n} \ln[\mathrm{E}_i^\pi\, \mathrm{e}^{\gamma \sum\limits_{k=0}^{n-1} \tilde{\varphi}_{x_k,x_{k+1}}(w^*,g^*)}] = 0, \tag{29}$$

$$\lim_{n\to\infty} \frac{1}{n} Z_i^\pi(\gamma,n) = \hat{g} \Longleftrightarrow \lim_{n\to\infty} \frac{1}{n} \ln[\mathrm{E}_i^\pi\, \mathrm{e}^{\gamma \sum\limits_{k=0}^{n-1} \tilde{\varphi}_{x_k,x_{k+1}}(\hat{w},\hat{g})}] = 0. \tag{30}$$

Proof. Recall that by (3),(4),(5) if the process starts in state $i \in \mathcal{I}$ and policy $\pi = (f^n)$ is followed the $\gamma$-sensitive average reward

$$J_i^\pi(\gamma) := \liminf_{n\to\infty} \frac{1}{n} Z_i^\pi(\gamma,n) \ \ \text{resp.} \ \ J_i^\pi(\gamma) := \limsup_{n\to\infty} \frac{1}{n} Z_i^\pi(\gamma,n)$$

where $Z_i^\pi(\gamma,n) = \frac{1}{\gamma} \ln U_i^\pi(\gamma,n)$, if the risk-sensitive minimal, resp. maximal, average reward is considered. From (19),(20) and Lemma 2.1 considered for policies minimizing or maximizing average reward it holds:

$$Z_i^\pi(\gamma,n) = n\hat{g} + \hat{w}_i + \hat{\Phi}_i^\pi(\gamma,n) + \hat{h}_i(n) \tag{31}$$

$$Z_i^\pi(\gamma,n) = ng^* + w_i^* + \Phi_i^{*,\pi}(\gamma,n) + h_i^*(n) \tag{32}$$

where $\quad |\hat{h}_i(n)| \le \max_i |\hat{w}_i|, \quad |h_i^*(n)| \le \max_i |w_i^*|$

$$\hat{\Phi}_i^\pi(\gamma,n) = \ln \mathrm{E}_i^\pi \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k,X_{k+1}}(\hat{w},\hat{g})}$$

$$\Phi_i^{*,\pi}(\gamma,n) = \ln \mathrm{E}_i^\pi \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k,X_{k+1}}(w^*,g^*)}.$$

Since $\mathrm{E}_i^\pi \mathrm{e}^{\gamma\tilde{\varphi}_{X_k,X_{k+1}}(\hat{w},\hat{g})} \ge 1$ for any policy $\pi = (f^n)$ and equals 1 if policy $\hat{\pi} \sim (\hat{f})$ is followed, we can conclude that $\hat{\Phi}_i^\pi(\gamma,n) \ge 1$ and equals 1 if stationary policy $\hat{\pi} \sim (\hat{f})$ is followed, i. e. $\ln \hat{\Phi}_i^\pi(\gamma,n) \ge 0$ and equal to 0 for $\pi \sim (\hat{f})$. Similarly, since $\mathrm{E}_i^\pi \mathrm{e}^{\gamma\tilde{\varphi}_{X_k,X_{k+1}}(w^*,g^*)} \le 1$ for any policy $\pi = (f^n)$, and equal to 1 if policy $\pi^* \sim (f^*)$ is followed. Then $\Phi_i^{*,\pi}(\gamma,n) \le 1$ and equals unity if stationary policy $\pi^* \sim (f^*)$ is followed, i. e. $\ln \mathrm{E}_i^\pi \mathrm{e}^{\gamma\tilde{\varphi}_{X_k,X_{k+1}}(w^*,g^*)} \le 0$ and for $\pi \sim (f^*)$ equal to 0. □

## 4. POISSONIAN EQUATIONS AND NONNEGATIVE MATRICES

In what follows we present necessary and sufficient conditions for the existence of a solution of optimality equations (26)–(28) for unichain models as well as for a very specific case of multi-chain models.

To this end, 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 $q_{ij}(f_i) := p_{ij}(f_i) \cdot \mathrm{e}^{\gamma r_{ij}}$ (26) can be alternatively written as the following set of equations

$$\rho(f) v_i(f) \quad = \quad \sum_{j\in\mathcal{I}} q_{ij}(f_i)\, v_j(f) \quad (i \in \mathcal{I}) \tag{33}$$

and (27), (28) can be rewritten as the following sets of nonlinear equations (here $\hat{v}_i := \mathrm{e}^{\gamma \hat{w}_i}$, $v^*_i := \mathrm{e}^{\gamma w^*_i}$, $\hat{\rho} = \mathrm{e}^{\gamma \hat{g}}$, $\rho^* := \mathrm{e}^{\gamma g^*}$)

$$\rho^* v^*_i = \max_{f\in\mathcal{F}} \sum_{j\in\mathcal{I}} q_{ij}(f_i)\, v^*_j, \quad \hat{\rho}\, \hat{v}_i = \min_{f\in\mathcal{F}} \sum_{j\in\mathcal{I}} q_{ij}(f_i)\, \hat{v}_j \quad (i \in \mathcal{I}) \tag{34}$$

called *$\gamma$-average reward/cost optimality equation in multiplicative form.*

For what follows it is convenient to consider (33), (34) in matrix form. To this end, on introducing the $N \times N$ matrix $Q(f) = [q_{ij}(f_i)]$ and (column) $N$-vector $v(f) = [v_i(f)]$, from (33) we get

$$\rho(f)\, v(f) = Q(f)\, v(f). \tag{35}$$

Since $Q(f)$ is a nonnegative matrix by the well-known Perron–Frobenius theorem (see, e. g. [14]) $\rho(f)$ is the spectral radius of $Q(f)$ that is equal to the maximum positive eigenvalue of $Q(f)$. Moreover, the corresponding right and left eigenvectors $v(f)$ and $z(f)$ ($N$-row vector), called the Perron eigenvectors, can be selected nonnegative. In particular, for the left (row) eigenvector $z(f)$ it holds:

$$\rho(f)\, z(f) = z(f)\, Q(f). \tag{36}$$

Moreover, if $Q(f)$ is irreducible the Perron eigenvectors can be selected strictly positive, i. e. (35), (36) hold with $v(f) > 0, z(f) > 0$.[3] Using matrix notations the symbol $I$ is reserved for identity matrix, $e$ denotes unit (column) vector.

Similarly, for $v(f^*) = v^*$, $v(\hat{f}) = \hat{v}$ (34) can be written in matrix form as

$$\rho^* v^* = \max_{f\in\mathcal{F}} Q(f) \cdot v^*, \qquad \hat{\rho}\, \hat{v} = \min_{f\in\mathcal{F}} Q(f) \cdot \hat{v}. \tag{37}$$

Recall that vectorial maximum and minimum in (37) should be considered componentwise and that $\hat{v}$, $v^*$ are unique up to multiplicative constant and strictly positive if $Q(\hat{f})$, $Q(f^*)$ is irreducible.

Moreover, strictly positive Perron eigenvectors still exist for reducible nonnegative matrices with specific structures. Since every reducible matrix can be written in a block-triangular form, an irreducible class of $Q(f)$ is called *basic* if and only if its spectral radius is equal to $\rho(f)$, else is *non-basic.*

[3] In vector inequality $a \geq b$ denotes that $a_i \geq b_i$ for all elements of the vectors $a$, $b$, and $a_i > b_i$ at least for one $i$, but not for all $i$'s, and $a > b$ if and only if and $a_i > b_i$ for all $i$'s.

Since (nonnegative) elements of $Q(f)$, denoted $q_{ij}(f_i) := p_{ij}(f_i) \cdot \mathrm{e}^{\gamma r_{ij}}$ are positive if and only if $p_{ij}(f_i)$ is positive, we say that some class of $Q(f)$ (or state $i \in \mathcal{I}$) *has access* to another class of $Q(f)$ (or state $j \in \mathcal{I}$) if and only if the same holds of the corresponding classes (or states) of the transition probability matrix $P(f)$. The same is used for communicating classes of $Q(f)$. The class that has no access to any other class is called *final*. Observe that in transition probability matrix recurrent classes are the basic and final classes. Recall that the spectral radius of any eigenvalue of a nonnegative matrix cannot be non-greater than the Perron eigenvalue.

Supposing that there exist strictly positive Perron eigenvectors the problem is easier and no partition of the state space is necessary. To this end we start our analysis by finding conditions guaranteeing existence of strictly positive Perron eigenvectors.

Necessary and sufficient condition for the existence of a strictly positive right eigenvector $v(f)$ of a nonnegative matrix $Q(f)$ with $f \in \mathcal{F}$ can be formulated as follows (see, e. g. [14]):

**Condition A.** If for suitable labelling of states of the underlying Markov chain (i. e. on suitably permuting rows and corresponding columns of $Q(f)$) it is possible to decompose $Q(f)$ such that:

$$Q(f) = \left[ \begin{array}{cc} Q_{(\mathrm{NN})}(f) & Q_{(\mathrm{NB})}(f) \\ 0 & Q_{(\mathrm{BB})}(f) \end{array} \right] \tag{38}$$

where $Q_{(\mathrm{NN})}(f)$ and $Q_{(\mathrm{BB})}(f)$ (with spectral radius $\rho_{(\mathrm{N})}(f)$ and $\rho_{(\mathrm{B})}(f)$) are (in general reducible) matrices such that:

- $\rho_{(\mathrm{N})}(f) < \rho(f)$, i. e. each irreducible class of $Q_{(\mathrm{NN})}(f)$ is non-basic,
- $\rho_{(\mathrm{B})}(f) = \rho(f)$ and $Q_{(\mathrm{BB})}(f)$ is block-diagonal, in particular,

$$Q_{(\mathrm{BB})}(f) = \left[ \begin{array}{ccc} Q_{11}(f) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Q_{rr}(f) \end{array} \right] \tag{39}$$

  where $Q_{(ii)}(f)$ (with $i = 1, \dots, r$) are irreducible submatrices with spectral radii $\rho_i(f) = \rho(f)$ with no access to any other class, i. e. $Q_{(ii)}(f)$ are final basic classes of $Q(f)$,

- each irreducible class of $Q_{(\mathrm{NN})}(f)$ *has access* to some basic class of $Q(f)$, hence at least some elements of $Q_{(\mathrm{NB})}(f)$ must be nonvanishing, in contrast to irreducible classes of $Q_{(\mathrm{BB})}(f)$ that are the *final classes* (having no access to any other class).

Using the matrix decomposition according to (38), the right Perron eigenvector $v(f)$ can be also decomposed as indicated in the next displayed formula

$$\rho(f) \left[ \begin{array}{c} v_{(\mathrm{N})}(f) \\ v_{(\mathrm{B})}(f) \end{array} \right] = \left[ \begin{array}{cc} Q_{(\mathrm{NN})}(f) & Q_{(\mathrm{NB})}(f) \\ 0 & Q_{(\mathrm{BB})}(f) \end{array} \right] \left[ \begin{array}{c} v_{(\mathrm{N})}(f) \\ v_{(\mathrm{B})}(f) \end{array} \right] \tag{40}$$

From (40) we conclude that $v_{(\mathrm{N})}(f) = [\rho(f)I - Q_{(\mathrm{NN})}(f)]^{-1}Q_{(\mathrm{NB})}(f)v_{(\mathrm{B})}(f)$. Similar decomposition can be also used for any column vector of an appropriate dimension.

To guarantee existence of strictly positive Perron eigenvector for any $Q(f)$ with $f \in \mathcal{F}$ we make

**Assumption 2.** There exists state $i_0 \in \mathcal{I}$ such that for every $Q(f)$ with $f \in \mathcal{F}$:
(i) $i_0$ belongs to the *basic* class of $Q(f)$ that is unique.
(ii) $i_0$ is *accessible* from any state $i \in \mathcal{I}$.

**Theorem 2.** Let Assumption 2 hold. Then the Poissonian equation (26) holds for any $f \in \mathcal{F}$. Minimal (resp. maximal) values $g(f)$, denoted $\hat{g}$ (resp. $g^*$), are the solution of (27) and (28).

Proof. (26) follows immediately from the Perron–Frobenius theorem. Using policy iterations we can also verify (27) and (28), see [25] for details. □

**Remarks.**

- If the transition probability matrix $P(f)$ is unichain (i.e. spectral radius is equal to 1 and the right Perron eigenvector is a unit vector) then the (unique) class of recurrent states is its basic class and each irreducible class of transient states is a nonbasic class (with spectral radius less than one) accessible to the recurrent class.

- Since elements $q_{ij}(f) = p_{ij}(f)\mathrm{e}^{\gamma r_{ij}}$, for $\gamma$ sufficiently close to null each transient class of $P(f)$ is a nonbasic class of $Q(f)$ (see also [7], Thm.3.1). Obviously, if Condition A holds and $\gamma \neq 0$ the spectral radius of each nonbasic class of $Q(f)$ must be less than $\rho(f)$.

- From (40) we immediately conclude that

$$\begin{aligned} v_{(\mathrm{N})}(f) &= [\rho(f)I - Q_{(\mathrm{NN})}(f)]^{-1}Q_{(\mathrm{NB})}(f)v_{(\mathrm{B})}(f) \\ &= (\rho(f))^{-1}\sum_{k=0}^{\infty}[(\rho(f))^{-1}Q_{(\mathrm{NN})}(f)]^{k}Q_{(\mathrm{NB})}(f)v_{(\mathrm{B})}(f) \end{aligned}$$

  hence the total reward earned from a transient state of $P(f)$ up to reaching a recurrent state of $P(f)$ is finite, and also the expected number of transitions up to reaching a recurrent state of $P(f)$ must be finite.

- Triangular structure of $Q_{(\mathrm{NN})}(f)$. In this case elements of the main diagonal of $Q_{(\mathrm{NN})}(f)$ can be considered as nonbasic classes of $Q(f)$ and condition A is fulfilled if and only if each of diagonal elements is less than $\rho(f)$. Observe that the spectral radius of $\rho_{(\mathrm{N})}(f)$ of $Q_{(\mathrm{NN})}(f) \neq 0$ equals null if and only all diagonal elements of $Q_{(\mathrm{NN})}(f)$ equal null. Then each transient state of $P(f) = [p_{ij}(f_i)]$ (i.e. each state of $Q_{(\mathrm{NN})}(f)$) is absorbed in the recurrent class of $P(f)$ after a finite number of transitions (at most equal to the number of transient states), see [6, 10].

- If Assumption 2 holds then $Q_{(\mathrm{BB})}(f)$ contains a single basic class. However, $Q(f)$ has a strictly positive eigenvector if $Q_{(\mathrm{BB})}(f)$ contains several basic classes with spectral radii equal to $\rho(f)$ and $\rho_{\mathrm{N}}(f) < \rho(f)$ (cf. (39),(40)). Of course, Perron eigenvalue of each basic class equals $\rho(f)$.

## 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. For the risk-sensitive case our analysis is mostly restricted to unichain models and some additional assumptions are made. Finally, it is indicated how suitable properties of unichain risk-sensitive models can be employed for the analysis of a very specific risk-sensitive multichain models.

Equations (26)–(27) were first analyzed by Howard and Matheson in [16] via the Perron–Frobenius theory of positive matrices under the condition that the state and action spaces are finite and that under every policy the whole state space is an aperiodic communicating class. Extension of this result can be found in [3], [7] showing that if the state space is a communicating class the solution of (26) is unique up to multiplicative constant. In contrast to the risk-neutral model, if the resulting transition probability matrix $P(f)$ is unichain and contains also transient states, solution of equations (26)–(28) can be guaranteed only for the small values of the risk sensitivity coefficient (see e. g. [7, 8, 9, 25]). Conditions guaranteeing existence of solutions to (26)–(27) were studied in many papers, see e. g. [3–12, 25–28].

In the present paper via the Perron–Frobenius theory of nonnegative matrices we presented necessary and sufficient conditions for existence of a solution to the set of equations (26)–(28) if the transition probability matrix $P(f)$ is unichain, and extended these result for a very specific class of multi-chain models.

ACKNOWLEDGEMENT

The research was supported by the Czech Science Foundation Republic under Grants 15-10331S, 18-02739S.

(Received December 8, 2017)

## REFERENCES

[1] A. Arapostathis, V. S. Borkar, F. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus: Discrete-time controlled Markov processes with average cost criterion: A survey. SIAM J. Control Optim. *31* (1993), 282–344. DOI:10.1137/0331018

[2] J. Bather: Optimal decisions procedures for finite Markov chains, Part II. Adv. Appl. Probab. *5* (1973), 328–339. DOI:10.2307/1426039

[3] T. D. Bielecki, D. Hernández-Hernández, and S. R. Pliska: Risk-sensitive control of finite state Markov chains in discrete time, with application to portfolio management. Math. Methods Oper. Res. *50* (1999), 167–188. DOI:10.1007/s001860050094

[4] R. Cavazos-Cadena: Value iteration and approximately optimal stationary policies in finite-state average Markov chains. Math. Methods Oper. Res. *56* (2002), 181–196. DOI:10.1007/s001860200205

[5] R. Cavazos-Cadena: 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), 2, 263–285. DOI:10.1007/s001860200256

[6] R. Cavazos-Cadena: Solution of the average cost optimality equation for finite Markov decision chains: risk-sensitive and risk-neutral criteria. Math. Methods Oper. Res. *70* (2009), 541–566. DOI:10.1007/s00186-008-0277-y

[7] R. Cavazos-Cadena and F. Fernandez-Gaucherand: Controlled Markov chains with risk-sensitive criteria: average cost, optimality equations and optimal solutions. Math. Methods Oper. Res. *43* (1999), 121–139.

[8] R. Cavazos-Cadena and D. Hernández-Hernández: A characterization exponential functionals in finite Markov chains. Math. Methods Oper. Res. *60* (2004), 399–414. DOI:10.1007/s001860400373

[9] R. Cavazos-Cadena and D. Hernández-Hernández: A characterization of the optimal risk-sensitive average cost in finite controlled Markov chains. Ann. Appl. Probab. *15* (2005), 175–212. DOI:10.1214/105051604000000585

[10] R. Cavazos-Cadena and D. Hernández-Hernández: Necessary and sufficient conditions for a solution to the risk-sensitive Poisson equation on a finite state space. System Control Lett. *58* (2009), 254–258. DOI:10.1016/j.sysconle.2008.11.001

[11] R. Cavazos-Cadena and R. Montes-de-Oca: The value iteration algorithm in risk-sensitive average Markov decision chains with finite state space. Math. Oper. Res. *28* (2003), 752–756. DOI:10.1287/moor.28.4.752.20515

[12] R. Cavazos-Cadena and R. Montes-de-Oca: Nonstationary value iteration in controlled Markov chains with risk-sensitive average criterion. J. Appl. Probab. *42* (2005), 905–918. DOI:10.1017/s0021900200000991

[13] R. Cavazos-Cadena, A. Feinberg, and R. Montes-de-Oca: A note on the existence of optimal policies in total reward dynamic programs with compact action sets. Math. Oper. Res. *25* (2000), 657–666. DOI:10.1287/moor.25.4.657.12112

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

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

[16] R. A. Howard and J. Matheson: Risk-sensitive Markov decision processes. Manag. Sci. *23* (1972), 356–369. DOI:10.1287/mnsc.18.7.356

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

[18] P. Mandl: Estimation and control in Markov chains. Adv. Appl. Probab. *6* (1974), 40–60. DOI:10.2307/1426206

[19] H. Markowitz: Portfolio selection. J. Finance *7* (1952), 77–92. DOI:10.1111/j.1540-6261.1952.tb01525.x

[20] H. Markowitz: Portfolio Selection – Efficient Diversification of Investments. Wiley, New York 1959.

[21] M. L. Puterman: Markov Decision Processes – Discrete Stochastic Dynamic Programming. Wiley, New York 1994. DOI:10.1002/9780470316887

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

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

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

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

[26] K. Sladký: Risk-sensitive and average optimality in Markov decision processes. In: Proc. 30th Int. Conf. Math. Meth. Economics 2012, Part II (J.Ramík and D.Stavárek, eds.), Silesian University, School of Business Administration, Karviná 2012, pp. 799–804. DOI:10.1007/3-540-32539-5_125

[27] K. Sladký: Risk-sensitive and mean variance optimality in Markov decision processes. Acta Oeconomica Pragensia *7* (2013), 146–161.

[28] N. M. van Dijk and K. Sladký: On the total reward variance for continuous-time Markov reward chains. J. Appl. Probab. *43* (2006), 1044–1052. DOI:10.1017/s0021900200002412

*Karel Sladký, Institute of Information Theory and Automation, The Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.*
*e-mail: sladky@utia.cas.cz*