Risk-Sensitive and Mean Variance Optimality in Markov Decision Processes

Karel Sladký, Milan Sitař

Department of Econometrics
Institute of Information Theory and Automation of the AS CR
Pod Vodárenskou věží 4, 182 08 Praha 8
sladky@utia.cas.cz, milan.sitar@tiscali.cz

Abstract

In this note, we compare two approaches for handling risk-variability features arising in discrete-time Markov decision processes: models with exponential utility functions and mean variance optimality models. Computational approaches for finding optimal decision with respect to the optimality criteria mentioned above are presented and analytical results showing connections between the above optimality criteria are discussed.

Keywords

Markov decision chains, exponential utility functions, certainty equivalent, expectation and variance of cumulative rewards, mean variance optimality, asymptotic behaviour

1 Introduction

The usual optimization criteria examined in the literature on optimization of Markov reward processes, e.g. total discounted or mean reward, may be quite insufficient to characterize the problem from the point of the decision maker. To this end it is necessary to select more sophisticated criteria that reflect also the variability-risk features of the problem.

Perhaps the best known approach how to handle such problems stems from the classical work of Markowitz [19] on mean variance selection rules for the portfolio selection problem. Following the mean variance selection rule, the investor selects from among a given set of investment alternatives only investments with a higher mean and lower variance than a member of the given set.

The mean variance selection rule can also be employed in Markovian decision models. Following this approach along with the total reward or long run average expected return (i.e. the mean reward per transition) we consider total or average variance of the (long run) cumulative rewards. For details see [11, 12, 14, 15, 16, 18, 29, 30], the review paper by White [32], and also recent results of the present authors [22, 24, 25, 26, 27]. It is important to notice that in many of the above papers the long run average "variance" is considered only with respect to one-stage reward variance and not to variance of cumulative rewards; hence it is more appropriate to speak about "average variability" instead of "average variance." As it was shown on a number of numerical examples in [22], optimal solutions based on "average variability" are mostly different of optimal solutions based on precisely calculated "average variance."

Another possible approach how to attack the variability-risk features arising in Markovian decision problems is to consider, instead of linear objective functions, exponential utility functions. Recall that exponential utility functions are the most widely used non-linear utility functions, cf. [6], and only linear and exponential functions are separable and hence appropriate for sequential decisions. Furthermore, Kirkwood [17] shows that in most cases an appropriately chosen exponential utility function is a very good approximation for general utility function. In [10] Howard demonstrates importance of exponential functions for treatment of a wide range of individual and risk preferences. The research of Markov decision processes with exponential objective functions, called risk-sensitive Markov decision processes, was initiated in the seminal paper by Howard and Matheson [9] and followed by many other researchers in recent years (see e.g. [3, 4, 5, 13, 23, 28]).

In this note we focus attention on risk-sensitive optimality criteria (i.e. the case when expectation of the stream of rewards generated by the Markov processes is evaluated by an exponential utility function) and their connections with mean-variance optimality (i.e. the case when a suitable combination of the expected total reward and its variance, usually considered per transition, is selected as a reasonable optimality criterion).

It is well known from the literature (see e.g. [31]) that for an exponential utility function, say $u^\gamma(\cdot)$, i.e. utility function with constant risk sensitivity $\gamma \in \mathbb{R}$, the utility assigned to the (random) reward ξ is given by

$$u^\gamma(\xi) := \begin{cases} \operatorname{sign}(\gamma)\, \exp(\gamma\xi) & \text{if } \gamma \neq 0 \\ \xi & \text{for } \gamma = 0. \end{cases} \tag{1.1}$$

Obviously $u^\gamma(\cdot)$ is continuous and strictly increasing. Moreover, if $\gamma > 0$ then $u^\gamma(\xi) = \exp(\gamma\xi)$ is convex and the decision maker is risk seeking. On the other hand if $\gamma < 0$ then $u^\gamma(\xi) = -\exp(\gamma\xi)$ is concave and the decision maker is risk averse.

The following facts are useful in the sequel:

1. For $U^{(\gamma)}(\xi) := \mathsf{E} \exp(\gamma\xi)$ the Taylor expansion around $\gamma = 0$ reads (in what follows E is reserved for expectation)

$$U^{(\gamma)}(\xi) = 1 + \mathsf{E} \sum_{k=1}^{\infty} \frac{(\gamma\xi)^k}{k!} = 1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} \cdot \mathsf{E}\, \xi^k. \tag{1.2}$$

Observe that in (1.2) the first (resp. second) term of the Taylor expansion is equal to $\gamma \mathsf{E}\,\xi$ (resp. $\frac{1}{2}(\gamma^2)\mathsf{E}\,\xi^2$). In particular, if for random variables ξ, ζ with $\mathsf{E}\,\xi = \mathsf{E}\,\zeta$ it holds $\mathsf{E}\,\xi^2 < \mathsf{E}\,\zeta^2$ (or equivalently $\operatorname{var}\xi < \operatorname{var}\zeta$) then there exists $\gamma_0 > 0$ such that $U^{(\gamma)}(\xi) < U^{(\gamma)}(\zeta)$ for any $\gamma \in (-\gamma_0, \gamma_0)$.

2. For $Z(\xi)$, the certainty equivalent of the (random) variable ξ, given by the condition $u^\gamma(Z(\xi)) = \mathsf{E}[u^\gamma(\xi)]$, we immediately get

$$Z(\xi) = \begin{cases} \frac{1}{\gamma} \ln\{\mathsf{E}\ [\exp(\gamma\xi)]\} & \text{if } \gamma \neq 0 \\ \mathsf{E}\,[\xi] & \text{for } \gamma = 0. \end{cases} \tag{1.3}$$

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

$$\begin{aligned} Z(\xi) &> \mathsf{E}\,\xi \qquad (\text{if } \gamma > 0, \ \text{the risk seeking case}) \\ Z(\xi) &< \mathsf{E}\,\xi \qquad (\text{if } \gamma < 0, \ \text{the risk averse case}) \\ Z(\xi) &= \mathsf{E}\,\xi \qquad (\text{if } \gamma = 0, \ \text{the risk neutral case}) \end{aligned}$$

3. Finally, recall that exponential utility function considered in (1.1) is separable what is very important for sequential decision problems, i.e. $u^\gamma(\xi^{(1)} + \xi^{(2)}) = \operatorname{sign}(\gamma)\, u^\gamma(\xi^{(1)}) \cdot u^\gamma(\xi^{(2)})$.

4. In economic models (see e.g. [1], [31]) we usually assume that the utility function $u(\cdot)$ is increasing (i.e. $u'(\cdot) > 0$), concave (i.e. $u''(\cdot) < 0$, what is fulfilled in (1.1) for $\gamma < 0$) with $u(0) = 0$ and $u'(0) < \infty$ (so called the Inada condition).

Since a linear transformation of the utility function $u^\gamma(\xi)$ preserves the original preferences (cf. [1],[31]) we shall also consider the utility functions

$$\begin{aligned} \bar{u}^\gamma(x) &= 1-\exp(\gamma x), \quad \text{where } \gamma<0 \quad \text{(the risk averse case)} &(1.4)\\ \tilde{u}^\gamma(x) &= \exp(\gamma x)-1, \quad \text{where } \gamma>0 \quad \text{(the risk seeking case)} &(1.5)\end{aligned}$$

and the function $\bar{u}^\gamma(x)$ satisfies all above conditions imposed on a utility function in economy theory. Observe that the Taylor expansions of $\bar{u}^\gamma(x)$ and of $\tilde{u}^\gamma(x)$ read

$$\bar{u}^\gamma(x) = \sum_{k=1}^{\infty}(-1)^{k+1}\frac{|\gamma|^k}{k!}\cdot x^k, \quad \text{where } \gamma<0, \quad \tilde{u}^\gamma(x) = \sum_{k=1}^{\infty}\frac{\gamma^k}{k!}\cdot x^k, \quad \text{where } \gamma>0 \tag{1.6}$$

and if $x=\xi$ is a random variable for the expected utilities we have

$$\bar{U}^\gamma(\xi) := \mathsf{E}\,\bar{u}^\gamma(\xi) = \sum_{k=1}^{\infty}(-1)^{k+1}\frac{|\gamma|^k}{k!}\cdot \mathsf{E}\,\xi^k, \qquad \tilde{U}^\gamma(\xi) := \mathsf{E}\,\tilde{u}^\gamma(\xi) = \sum_{k=1}^{\infty}\frac{\gamma^k}{k!}\cdot \mathsf{E}\,\xi^k. \tag{1.7}$$

In this note we focus attention on properties of the expected utility and the corresponding certainty equivalents if the stream of obtained rewards is evaluated by exponential utility functions and their connections with more classical mean-variance optimality criteria.

2 Notation and Preliminaries

Consider a Markov decision chain $X=\{X_n,\ n=0,1,\ldots\}$ with finite state space $\mathcal{I}=\{1,\ldots,N\}$, finite set $\mathcal{A}_i=\{1,2,\ldots,K_i\}$ of possible decisions (actions) in state $i\in\mathcal{I}$ and the following transition and reward structure (we assume that in state i action $a\in\mathcal{A}_i$ is selected):

p_{ij}^a : transition probability from $i\to j$ $(i,j\in\mathcal{I})$,

r_{ij} : one-stage reward for a transition from $i\to j$,

r_i^a : expected value of the one-stage rewards incurred in state i,

$r_i^{(2),a}$: second moment of the one-stage rewards incurred in state i.

Obviously, $r_i^a=\sum_{j\in\mathcal{I}}p_{ij}^a\cdot r_{ij}$, $r_i^{(2),a}=\sum_{j\in\mathcal{I}}p_{ij}^a\cdot[r_{ij}]^2$ and hence the corresponding one-stage reward variance $\sigma_i^{2,a}=r_i^{(2),a}-[r_i^a]^2$.

Policy controlling the chain is a rule how to select actions in each state. In this note, we restrict on stationary policies, i.e. the rules selecting actions only with respect to the current state of the Markov chain X. Then a policy, say π, is determined by some decision vector f whose ith element $f_i\in\mathcal{A}_i$ identifies the action taken if the chain X is in state $X_n=i$; hence also the transition probability matrix $\boldsymbol{P}(f)$ of the Markov decision chain. Observe that the ith row of $\boldsymbol{P}(f)$ has elements $p_{i1}^{f_i},\ldots,p_{iN}^{f_i}$ and that $\boldsymbol{P}^*(f)=\lim_{n\to\infty}n^{-1}\sum_{k=0}^{n-1}[\boldsymbol{P}(f)]^k$ exists. In what follows, $\boldsymbol{R}=[r_{ij}]$ is the transition reward matrix, i.e. $\boldsymbol{R}$ is an $N\times N$ matrix of one-stage rewards. Similarly, $\boldsymbol{r}(f)$ is the (column) vector of one-stage expected rewards with elements $r_1^{f_1},\ldots,r_N^{f_N}$.

Let elements of the vectors $\boldsymbol{R}^\pi(n)$, $\boldsymbol{S}^\pi(n)$ and $\boldsymbol{V}^\pi(n)$ denote the first moment, the second moment and the variance of the (random) total reward $\xi_i^{(n)}(\pi)$ respectively received in the n next transitions of the considered Markov chain X if policy $\pi\sim(f)$ is followed, given the initial state $X_0=i$. In what follows we sometimes abbreviate $\xi_i^{(n)}(\pi)$ by $\xi_i^{(n)}$ or by $\xi^{(n)}$ if the dependence on policy $\pi\sim(f)$ or initial state $X_0=i$ is obvious.

More precisely, for the elements of $\boldsymbol{R}^\pi(n)$, $\boldsymbol{S}^\pi(n)$ and $\boldsymbol{V}^\pi(n)$ we have

$$R_i^\pi(n)=\mathsf{E}_i^\pi[\xi^{(n)}], \quad S_i^\pi(n)=\mathsf{E}_i^\pi[\xi^{(n)}]^2, \quad V_i^\pi(n)=\boldsymbol{\sigma}_i^{2,\pi}[\xi^{(n)}]$$

where $\xi^{(n)} = \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}$ and E_i^π, $\boldsymbol{\sigma}_i^{2,\pi}$ are standard symbols for expectation and variance if policy π is selected and $X_0 = i$. Moreover, if $m < n$ we can write $\xi_{X_0}^{(n)} = \xi_{X_0}^{(m)} + \xi_{X_m}^{(m,n)}$, where $\xi_{X_m}^{(m,n)} = \sum_{k=m}^{n-1} r_{X_k,X_{k+1}}$ is reserved for the (random) reward obtained from the mth up to the nth transition.

Recall that

$$R_i^\pi(n+1) \;=\; r_i^{f_i} + \sum_{j\in\mathcal{I}} p_{ij}^{f_i} R_j^\pi(n) \tag{2.1}$$

or in vector notation

$$\boldsymbol{R}^\pi(n+1) \;=\; \boldsymbol{r}(f) + \boldsymbol{P}(f)\cdot\boldsymbol{R}^\pi(n). \tag{2.2}$$

Similarly, if the chain starts in state i and policy $\pi \sim (f)$ is followed then from (1.2), (1.3) for $\xi = \xi^{(n)}$ for the expected utility $U_i^\pi(\gamma,n)$, the certainty equivalent $Z_i^\pi(\gamma,n)$ and its mean value $J_i^\pi(\gamma,n)$ we have

$$U_i^\pi(\gamma,n) \;:=\; \mathsf{E}_i^\pi[\exp(\gamma\xi^{(n)})] = \mathsf{E}_i^\pi \exp[\gamma\,(r_{i,X_1} + \xi_{X_1}^{(1,n)})], \tag{2.3}$$

$$Z_i^\pi(\gamma,n) \;:=\; \frac{1}{\gamma}\,\ln\{\mathsf{E}_i^\pi\,[\exp(\gamma\xi^{(n)})]\} \qquad \text{for } \gamma\neq 0, \tag{2.4}$$

$$J_i^\pi(\gamma,n) \;:=\; \lim_{n\to\infty}\frac{1}{n} Z_i^\pi(\gamma,n) \tag{2.5}$$

and hence for the expectation of the utility functions $\bar{u}^\gamma(\xi^{(n)})$ and $\tilde{u}^\gamma(\xi^{(n)})$ we have (cf. (1.7))

$$\bar{U}_i^\pi(\gamma,n) := 1 - U_i^\pi(\gamma,n), \qquad \tilde{U}_i^\pi(\gamma,n) := U_i^\pi(\gamma,n) - 1. \tag{2.6}$$

In what follows let $\boldsymbol{U}^\pi(\gamma,n)$, resp. $\boldsymbol{Z}^\pi(\gamma,n)$, be the vector of expected utilities, resp. certainty equivalents, with elements $U_i^\pi(\gamma,n)$, resp. $Z_i^\pi(\gamma,n)$.

Conditioning in (2.3) on X_1, since policy $\pi \sim (f)$ is stationary, from (2.3) we immediately get the recurrence formula

$$U_i^\pi(\gamma,n+1) \;=\; \sum_{j\in\mathcal{I}} p_{ij}^{f_i}\cdot \mathrm{e}^{\gamma r_{ij}}\cdot U_j^\pi(\gamma,n) = \sum_{j\in\mathcal{I}} q_{ij}^{f_i}\cdot U_j^\pi(\gamma,n) \quad \text{with } U_i^\pi(\gamma,0)=1 \text{ or} \tag{2.7}$$

in vector notation

$$\boldsymbol{U}^\pi(\gamma,n+1) \;=\; \boldsymbol{Q}(f)\cdot\boldsymbol{U}^\pi(\gamma,n) \qquad \text{with } \boldsymbol{U}^\pi(\gamma,n) = \boldsymbol{e}, \tag{2.8}$$

where $\boldsymbol{Q}(f) = [q_{ij}^{f_i}]$ with $q_{ij}^{f_i} := p_{ij}^{f_i}\cdot\mathrm{e}^{\gamma r_{ij}}$.

Observe that $\boldsymbol{Q}(f)$ is a nonnegative matrix, and by the Perron–Frobenius theorem (cf. [7]) the spectral radius $\rho(f)$ of $\boldsymbol{Q}(f)$ is equal to the maximum positive eigenvalue of $\boldsymbol{Q}(f)$. Moreover, if $\boldsymbol{Q}(f)$ is irreducible (i.e. if and only if $\boldsymbol{P}(f)$ is irreducible) the corresponding (right) eigenvector $\boldsymbol{v}(f)$ can be selected strictly positive, i.e.

$$\rho(f)\,\boldsymbol{v}(f) = \boldsymbol{Q}(f)\cdot\boldsymbol{v}(f) \qquad \text{with} \quad \boldsymbol{v}(f) > 0. \tag{2.9}$$

Moreover, under the above irreducibility condition it can be shown (cf. e.g. [9], [28]) that there exists decision vector $f^* \in \mathcal{A}$ such that

$$\boldsymbol{Q}(f)\cdot\boldsymbol{v}(f^*) \;\le\; \rho(f^*)\,\boldsymbol{v}(f^*) = \boldsymbol{Q}(f^*)\cdot\boldsymbol{v}(f^*), \tag{2.10}$$

$$\rho(f) \;\le\; \rho(f^*) \equiv \rho^* \qquad \text{for all } f\in\mathcal{A}. \tag{2.11}$$

In words, $\rho(f^*) \equiv \rho^*$ is the maximum possible eigenvalue of $\boldsymbol{Q}(f)$ over all $f \in \mathcal{A}$.

Throughout this note we make the following assumptions.

AS 1. For any stationary policy $\pi \sim (f)$, the transition probability matrix $\boldsymbol{P}(f)$ is irreducible (i.e. all states are communicating) and aperiodic, i.e. $\boldsymbol{P}(f)$ is ergodic (all states are recurrent and aperiodic).

Observe that under AS 1 the matrix $\boldsymbol{Q}(f)$ is irreducible for any $f \in \mathcal{A}$.

AS 2. Transition rewards are nonnegative and nonvanishing, i.e. $r_{ij} \ge 0$ for all $i,j \in \mathcal{I}$ and a strict inequality holds at least for one pair i,j.

Observe that under AS 2 all one-stage expected rewards $r_i(\cdot)$ are nonnegative.

3 Reward Variance and Expected Utility

Since for any integers $m < n$ $[\xi^{(n)}]^2 = [\xi^{(m)}]^2 + 2 \cdot \xi^{(m)} \cdot \xi^{(m,n)} + [\xi^{(m,n)}]^2$ we get

$$\mathsf{E}_i^\pi[\xi^{(n)}]^2 = \mathsf{E}_i^\pi[\xi^{(m)}]^2 + 2 \cdot \mathsf{E}_i^\pi[\xi^{(m)} \cdot \xi^{(m,n)}] + \mathsf{E}_i^\pi[\xi^{(m,n)}]^2. \tag{3.1}$$

In particular, for $m = 1$, $n := n+1$ if policy $\pi \sim (f)$ is followed we get for the second moment of the random reward $\xi^{(n)}$:

$$S_i^\pi(n+1) = \sum_{j\in\mathcal{I}} p_{ij}^{f_i} \cdot \{[r_{ij}]^2 + 2 \cdot r_{ij} \cdot R_j^\pi(n)\} + \sum_{j\in\mathcal{I}} p_{ij}^{f_i} \cdot S_j^\pi(n). \tag{3.2}$$

Since the variance $V_i(\cdot) = S_i(\cdot) - [R_i(\cdot)]^2$ from (3.2) we arrive at

$$V_i^\pi(n+1) = \sum_{j\in\mathcal{I}} p_{ij}^{f_i} \cdot \{[r_{ij} + R_j^\pi(n)]^2\} - [R_i^\pi(n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}^{f_i} \cdot V_j^\pi(n). \tag{3.3}$$

From the literature (see e.g. [8, 20, 21] it is well known that under AS 1 there exist vector $\boldsymbol{w}^\pi$ (with elements w_j^π), constant vector $\boldsymbol{g}^\pi$ and vector $\boldsymbol{\varepsilon}(n)$ (where all elements of $\boldsymbol{\varepsilon}(n)$ converge to zero geometrically) such that

$$\boldsymbol{R}^\pi(n) = \boldsymbol{g}^\pi \cdot n + \boldsymbol{w}^\pi + \boldsymbol{\varepsilon}(n) \Rightarrow \lim_{n\to\infty} n^{-1}\boldsymbol{R}^\pi(n) = \boldsymbol{g}^\pi = \boldsymbol{P}^*(f) \cdot \boldsymbol{r}(f). \tag{3.4}$$

The constant vector $\boldsymbol{g}^\pi$ (with elements g^π) along with vector $\boldsymbol{w}^\pi$ are uniquely determined by

$$\boldsymbol{w}^\pi + \boldsymbol{g}^\pi \ = \ \boldsymbol{r}(f) + \boldsymbol{P}(f) \cdot \boldsymbol{w}^\pi, \quad \boldsymbol{P}^*(f) \cdot \boldsymbol{w}^\pi = \boldsymbol{0}. \tag{3.5}$$

By using relations (3.3), (3.4) and (3.5) in a number of steps we arrive at (for details see [24, 26, 25, 27]):

$$\boldsymbol{V}^\pi(n+1) = \boldsymbol{s}(\pi) + \boldsymbol{P}(f) \cdot \boldsymbol{V}^\pi(n) + \boldsymbol{\varepsilon}^{(1)}(n) \tag{3.6}$$

where for elements of the vector $\boldsymbol{s}(\pi)$ we have

$$s_i(\pi) = \sum_{j\in\mathcal{I}} p_{ij}^{f_i} \cdot \{[r_{ij} + w_j^\pi]^2\} - [g^\pi + w_i^\pi]^2 \tag{3.7}$$

$$= \sum_{j\in\mathcal{I}} p_{ij}^{f_i} \cdot \{[r_{ij} - g^\pi + w_j^\pi]^2\} - [w_i^\pi]^2 \tag{3.8}$$

and elements of the vector $\boldsymbol{\varepsilon}^{(1)}(n)$ converge to zero geometrically.

In analogy with (3.4), (3.5) we can conclude that there exists vector $\boldsymbol{w}^{(2),\pi}$ along with a constant vector $\boldsymbol{g}^{(2),\pi}$ uniquely determined by

$$\boldsymbol{w}^{(2),\pi} + \boldsymbol{g}^{(2),\pi} = \boldsymbol{s}(\pi) + \boldsymbol{P}(f) \cdot \boldsymbol{w}^{(2),\pi}, \quad \boldsymbol{P}^*(f) \cdot \boldsymbol{w}^{(2),\pi} = \boldsymbol{0} \tag{3.9}$$

such that

$$\boldsymbol{V}^\pi(n) = \boldsymbol{g}^{(2),\pi} \cdot n + \boldsymbol{w}^{(2),\pi} + \boldsymbol{\varepsilon}(n) \implies \boldsymbol{g}^{(2),\pi} = \lim_{n\to\infty} \frac{\boldsymbol{V}^\pi(n)}{n} = \boldsymbol{P}^*(f) \cdot \boldsymbol{s}(\pi). \tag{3.10}$$

Moreover, from (3.5), (3.6), (3.10) we can also conclude after some algebra (observe that $\boldsymbol{P}^*(f) \cdot \boldsymbol{P}(f) \cdot \boldsymbol{w}^{(2),\pi} = \boldsymbol{P}^*(f) \cdot \boldsymbol{w}^{(2),\pi}$, cf. [24] for details) that

$$\boldsymbol{g}^{(2),\pi} = \bar{\boldsymbol{g}}^{(2),\pi} + 2 \cdot \boldsymbol{P}^*(f) \cdot \tilde{\boldsymbol{r}}(f,\pi) \tag{3.11}$$

where ($[\cdot]_{\text{sq}}$ denotes that elements of the vector are squared)

$\boldsymbol{r}^{(2)}(f)$ is a column vector with elements $r_i^{(2),f_i} = \sum_{j\in\mathcal{I}} p_{ij}^{f_i} \cdot [r_{ij}]^2$,

$\tilde{\boldsymbol{r}}(f,\pi)$ is a column vector with elements $\sum_{j\in\mathcal{I}} p_{ij}^{f_i}\cdot r_{ij}\cdot w_j^{\pi}$,

$[\boldsymbol{g}^{\pi}]_{\rm sq}$ is a constant vector with elements $[g^{\pi}]^2$,

$\tilde{\boldsymbol{g}}^{(2),\pi} = \boldsymbol{P}^*(f)\cdot \boldsymbol{r}^{(2)}(f)$ is a constant vector with elements $\tilde{g}^{(2),\pi}$, and

$\bar{\boldsymbol{g}}^{(2),\pi} = \tilde{\boldsymbol{g}}^{(2),\pi} - [\boldsymbol{g}^{\pi}]_{\rm sq}$ is a constant vector with elements $\bar{g}^{(2),\pi}$.

Obviously, $\tilde{g}^{(2),\pi}$ averages expected values of the second moments of one-stage rewards, $\bar{g}^{(2),\pi}$ denotes the average "one-stage reward variance" considered with respect to the mean reward g^{π} instead of the one-stage expected reward $r_i^{f_i}$ in state $i\in\mathcal{I}$, and the last term in (3.11) expresses the Markov dependence that occurs if the total variance of cumulative rewards is considered.

On the other hand, on iterating (2.7) we get if (stationary) policy $\pi\sim(f^*)$ is followed

$$\boldsymbol{U}^{\pi}(\gamma,n) = (\boldsymbol{Q}(f^*))^n\cdot \boldsymbol{e}. \tag{3.12}$$

Since under AS 1 the Perron eigenvector $\boldsymbol{v}(f^*)$ is strictly positive, there exist numbers $\alpha_1<\alpha_2$ such that $\alpha_1\cdot\boldsymbol{v}(f^*)\le \boldsymbol{e}\le \alpha_2\cdot\boldsymbol{v}(f^*)$ and hence

$$\alpha_1\cdot(\rho(f^*))^n\cdot\boldsymbol{v}(f^*)\le \boldsymbol{U}^{\pi}(\gamma,n)\le \alpha_2\cdot(\rho(f^*))^n\cdot\boldsymbol{v}(f^*). \tag{3.13}$$

From (3.13) we can see that the asymptotic behaviour of $\boldsymbol{U}^{\pi}(\gamma,n)$ heavily depends on $\rho(f^*)$, and the growth rate of each $U_i^{\pi}(\gamma,n)$ is the same and equal to $\rho(f^*)$.

On examining $\boldsymbol{Q}(f)$ we can easily conclude that:

If $\gamma<0$ then $\rho(f^*)<1$ for all $f\in\mathcal{A}$, hence $\lim_{n\to\infty}[\boldsymbol{Q}(f)]^n=\boldsymbol{0}$ and $U_i^{\pi}(\gamma,n)\to 0$ for all $i\in\mathcal{I}$ as $n\to\infty$ and the convergence is geometrical.

For $\gamma=0$ we have $\boldsymbol{Q}(f)=\boldsymbol{P}(f)$, the spectral radius of $\boldsymbol{Q}(f)$ equals one, and the corresponding right eigenvector $\boldsymbol{v}(f^*)$ is a constant vector. Then $\boldsymbol{U}^{\pi}(\gamma,n)\to\boldsymbol{P}^*(f)\cdot\boldsymbol{e}$, a constant vector.

If $\gamma>0$ then $\rho(f^*)>1$ for $f^*\in\mathcal{A}$, hence elements of $[\boldsymbol{Q}(f^*)]^n$ go to infinity and also $U_i^{\pi^*}(\gamma,n)\to\infty$ for all $i\in\mathcal{I}$ as $n\to\infty$.

Moreover, by (1.2) we can expect that for γ sufficiently close to null the growth of $U_i^{\pi}(\gamma,n)$ will be dominated by the first and second moment of $\xi^{(n)}$ occurring in (3.4).

Illustrative Example.

Consider a controlled Markov reward chain with 5 states and only three possible actions in state 1 (in the remaining states no option is possible). Hence only three transition probability matrices, say $\boldsymbol{P}(f^{(1)})$, $\boldsymbol{P}(f^{(2)})$, $\boldsymbol{P}(f^{(3)})$, are available that along with the reward matrix $\boldsymbol{R}$ fully characterize the transition and reward structures of the considered Markov reward chain. Observe that stationary policies $\pi^{(2)}\sim(f^{(2)})$, $\pi^{(3)}\sim(f^{(3)})$ identify a constant sequence of one-stage rewards. In particular, we have

$$\boldsymbol{P}(f^{(1)}) = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0\\ 0.5 & 0 & 0.5 & 0 & 0\\ 0.5 & 0.5 & 0 & 0 & 0\\ 0.5 & 0 & 0 & 0.5 & 0\\ 0.5 & 0 & 0 & 0 & 0.5 \end{bmatrix}, \qquad \boldsymbol{R} = \begin{bmatrix} 0 & 1 & 0 & 0.5 & 0.48\\ 0 & 0 & 1 & 0 & 0\\ 1 & 0 & 0 & 0 & 0\\ 0.5 & 0 & 0 & 0.5 & 0\\ 0.48 & 0 & 0 & 0 & 0.48 \end{bmatrix},$$

$$\boldsymbol{P}(f^{(2)}) = \begin{bmatrix} 0 & 0 & 0 & 1 & 0\\ 0.5 & 0 & 0.5 & 0 & 0\\ 0.5 & 0.5 & 0 & 0 & 0\\ 0.5 & 0 & 0 & 0.5 & 0\\ 0.5 & 0 & 0 & 0 & 0.5 \end{bmatrix}, \qquad \boldsymbol{P}(f^{(3)}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 1\\ 0.5 & 0 & 0.5 & 0 & 0\\ 0.5 & 0.5 & 0 & 0 & 0\\ 0.5 & 0 & 0 & 0.5 & 0\\ 0.5 & 0 & 0 & 0 & 0.5 \end{bmatrix}.$$

Obviously, if the chain starts in state 1 and action 2 (resp. 3) is selected then $r_{X_k,X_{k+1}}\equiv 0.5$ (resp. $r_{X_k,X_{k+1}}\equiv 0.48$) for all $k=0,1,\ldots$ and the chain visits only the states $1,4$ (resp. $1,5$). Hence for the expected reward, the second moment and the variance respectively, we have

$R_1(k) = 0.5k$, $S_1(k) = (0.5k)^2$, $V_1(k) \equiv 0$ (resp. $R_1(k) = 0.48k$, $S_1(k) = (0.48k)^2$, $V_1(k) \equiv 0$). On the contrary if the chain starts in state 1 and action 1 is selected on inspecting the chain we can see that the chain visits only the states $1, 2, 3$ and that the sequence of received rewards obeys a binomial distribution with parameter $p = 0.5$, and hence again $R_1(k) = 0.5k$, however $V_1(k) = k\,0.5(1 - 0.5) = 0.25k$.

The same results can be obtained by using the general formulas for expected reward and variance of the Markov reward chains as it is shown in the further text.

Observe that by (3.4), (3.5) for

$$\begin{aligned}
&\pi^{(1)} \sim (f^{(1)}), \quad g^\pi = 0.5, \quad \tilde{g}^{(2),\pi} = 0.5, \quad &&\text{hence } g^{(2),\pi} = 0.25,\\
&\pi^{(2)} \sim (f^{(2)}), \quad g^\pi = 0.5, \quad \tilde{g}^{(2),\pi} = 0.25, \quad &&\text{hence } g^{(2),\pi} = 0,\\
&\pi^{(3)} \sim (f^{(3)}), \quad g^\pi = 0.48, \quad \tilde{g}^{(2),\pi} = (0.48)^2, \quad &&\text{hence } g^{(2),\pi} = 0.
\end{aligned}$$

Of course, following policy $\pi^{(2)} \sim (f^{(2)})$ we get maximum possible mean reward and null variance, this policy is the best choice. However, the second best policy can be either $\pi^{(3)} \sim (f^{(3)})$ guaranteeing mean reward slightly less than then maximum one and the null variance or policy $\pi^{(1)} \sim (f^{(1)})$ giving the maximum mean reward, but the variance comparable with the mean reward. Considering optimality in accordance the weighted optimality criterion $\alpha g^\pi - (1-\alpha)g^{(2),\pi}$ (where $\alpha \in [0,1]$) it depends on the decision maker option how to selected the weighting coefficient α and prefer either policy $\pi^{(1)} \sim (f^{(1)})$ or policy $\pi^{(3)} \sim (f^{(3)})$. Since for $\alpha_0 = \frac{25}{27}$ it holds $\alpha_0\, g^{\pi_1} - (1-\alpha)\, g^{(2),\pi_1} = \alpha_0\, g^{\pi_3}$, if $\alpha \in [\alpha_0, 1]$ we prefer policy $\pi^{(1)}$ above policy $\pi^{(3)}$; if $\alpha < \alpha_0$ we consider policy $\pi^{(3)}$ as the second best.

Employing the risk-sensitive model and supposing that the chain starts in state 1 it is sufficient to consider the following matrices:

$$\widetilde{\boldsymbol{Q}}(f^{(1)}) = \begin{bmatrix} 0 & \frac{1}{2}\cdot \mathrm{e}^\gamma & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2}\cdot \mathrm{e}^\gamma \\ \frac{1}{2}\cdot \mathrm{e}^\gamma & \frac{1}{2} & 0 \end{bmatrix}, \quad \widetilde{\boldsymbol{Q}}(f^{(2)}) = \begin{bmatrix} 0 & \mathrm{e}^{\gamma 0.5} \\ \frac{1}{2}\mathrm{e}^{\gamma 0.5} & \frac{1}{2}\mathrm{e}^{\gamma 0.5} \end{bmatrix}, \quad \widetilde{\boldsymbol{Q}}(f^{(3)}) = \begin{bmatrix} 0 & \mathrm{e}^{\gamma 0.48} \\ \frac{1}{2}\mathrm{e}^{\gamma 0.48} & \frac{1}{2}\mathrm{e}^{\gamma 0.48} \end{bmatrix}$$

The right Perron eigenvector of each above matrix is a unit vector of appropriate dimension and for the spectral radii we get $\tilde{\rho}(f^{(1)}) = 0.5\,(\mathrm{e}^\gamma + 1)$, $\tilde{\rho}(f^{(2)}) = \mathrm{e}^{\gamma\,0.5}$, $\tilde{\rho}(f^{(3)}) = \mathrm{e}^{\gamma\,0.48}$. Taking into account maximum growth rate of the exponential utility function (1.7), and hence the maximal growth rate of the expected utility $\bar{U}^\gamma(\xi) = 1 - U^{(\gamma)}(\xi)$ (since the considered risk aversion coefficient γ is negative), obtained for the minimum possible Perron eigenvalue of the matrix $\widetilde{\boldsymbol{Q}}(\cdot)$, we get that $\tilde{\rho}(f^{(2)}) < \tilde{\rho}(f^{(3)})$ for $\gamma < 0$ implying that policy $\pi^{(2)} \sim (f^{(2)})$ is "better" than $\pi^{(3)} \sim (f^{(3)})$ with respect to the considered risk-sensitive criterion. Similarly, on comparing $\tilde{\rho}(f^{(3)}) = \mathrm{e}^{\gamma\,0.48}$ and $\tilde{\rho}(f^{(1)}) = 0.5\,(\mathrm{e}^\gamma + 1)$ we can decide whether $\pi^{(3)} \sim (f^{(3)})$ or $\pi^{(1)} \sim (f^{(1)})$ is the second best policy for the considered the value of the risk aversion coefficient γ.

Acknowledgement. This research was supported by the Czech Science Foundation under Grants 402/08/0107 and 402/07/1113.

References

[1] J. von Neumann and O. Morgenstern: Theory of Games and Economic Behaviour. Third Edition, Princeton Univ. Press, Princeton, NJ 1953.

[2] Ch. Barz: Risk-Averse Capacity Control in Revenue Management. Springer, Berlin–Heidelberg 2007.

[3] T. 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.

[4] 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.

[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), 253–285.

[6] J. L. Corner and P. D. Corner: Characterization of decision in decision analysis practice. J. Oper. Res. Soc. *46* (1995), 304–314.

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

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

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

[10] R. A. Howard: Decision analysis: Practice and promise. Manag. Sci. *34* (1988), 679–695.

[11] S. C. Jaquette: Markov decision processes with a new optimality criterion: small interest rates. Ann. Math. Statist. *43* (1972), 1894–1901.

[12] S. C. Jaquette: Markov decision processes with a new optimality criterion: discrete time. Ann. Statist. *1* (1973), 496–505.

[13] S. C. Jaquette: A utility criterion for Markov decision processes. Manag. Sci. *23* (1976), 43–49.

[14] J. Filar, L. C. M. Kallenberg, and H.-M. Lee: Variance penalized Markov decision processes. Mathem. Oper. Research *14* (1989), 147–161.

[15] Ying Huang and L. C. M. Kallenberg: On finding optimal policies for Markov decision chains: a unifying framework for mean-variance-tradeoffs. Mathem. Oper. Research *19* (1994), 434–448.

[16] H. Kawai: A variance minimization problem for a Markov decision process. European J. Oper. Research *31* (1987), 140–145. Springer, Berlin 2004, pp. 43–66.

[17] C. W. Kirkwood: Approximating risk aversion in decision analysis applications. Decision Analysis *1* (2004), 51–67.

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

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

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

[21] S. M. Ross: Applied Probability Models with Optimization Applications. Holden–Day, San Francisco, Calif. 1970.

[22] M. Sitař: Mean-Variance Optimality in Markov Decision Processes. Doctoral Thesis, Charles University, Prague 2006.

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

[24] K. Sladký and M. Sitař: Optimal solutions for undiscounted variance penalized Markov decision chains. In: Dynamic Stochastic Optimization (LNEMS 532, K. Marti, Y. Ermoliev, and G. Pflug, eds.), Springer, Berlin 2004, pp. 43–66.

[25] K. Sladký and M. Sitař: Algorithmic procedures for mean variance optimality in Markov decision chains. In: Operations Research Proceedings 2005 Springer, Berlin 2006, pp. 799–804.

[26] K. Sladký and M. Sitař: Mean–variance optimality in Markov decision chains. In: Mathematical Methods in Economics 2005, Gaudeamus, Hradec Králové 2005.

[27] K. Sladký: On mean reward variance in semi-Markov processes. Math. Methods Oper. Res. *62* (2005), 387–397.

[28] K. Sladký: Growth rates and average optimality in risk-sensitive Markov decision chains. Kybernetika *44* (2008), 206–217.

[29] M. J. Sobel: The variance of discounted Markov decision processes. J. Appl. Probab. *19* (1982), 794–802.

[30] M. J. Sobel: Maximal mean/standard deviation ratio in an undiscounted MDP. Oper. Research Lett. *4* (1985), 157–159.

[31] A. Takayama: Analytical Methods in Economics. Harvester Wheatsheaf, Hertfordshire 1994.

[32] D. J. White: Mean, variance and probability criteria in finite Markov decision processes: A review. J. Optim. Theory Appl. *6* (1988), 1–29.