Risk-sensitive and Mean Variance Optimality in Continuous-time Markov Decision Chains

Karel Sladký[1]

Abstract. In this note we consider continuous-time Markov decision processes with finite state and actions spaces where the stream of rewards generated by the Markov processes is evaluated by an exponential utility function with a given risk sensitivity coefficient (so-called risk-sensitive models). If the risk sensitivity coefficient equals zero (risk-neutral case) we arrive at a standard Markov decision process. Then we can easily obtain necessary and sufficient mean reward optimality conditions and the variability can be evaluated by the mean variance of total expected rewards. For the risk-sensitive case, i.e. if the risk-sensitivity coefficient is non-zero, for a given value of the risk-sensitivity coefficient we establish necessary and sufficient optimality conditions for maximal (or minimal) growth rate of expectation of the exponential utility function, along with mean value of the corresponding certainty equivalent. Recall that in this case along with the total reward also its higher moments are taken into account.

Keywords: Continuous-time Markov decision chains, exponential utility functions, certainty equivalent, mean-variance optimality, connections between risk-sensitive and risk-neutral models

JEL classification: C44, C61
AMS classification: 90C40, 60J10

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 variability-risk features of the problem. Perhaps the best known approaches stem from the classical work of Markowitz (cf. [9]) on mean variance selection rules, i.e. we optimize the weighted sum of average or total reward and its variance.

In this paper we restrict attention on continuous-time Markov decision chains models with finite state spaces where the variability is measured *either* by an exponential utility function with a given value of the risk-sensitivity coefficient *or* in the class of long-run average optimal policies by choosing policies with minimal mean-variance. Observe that in the former case along with the total reward also its higher moments are taken into account.

The article is structured as follows. Section 2 contains notations and summary of basic facts on continuous-time Markov reward processes. Markov models with exponential utility function (called risk-sensitive Markov chains) are studied in section 3 along with the corresponding moment generating functions. Section 4 is devoted to mean-variance optimality in continuous-time Markov decision chains; the results are mostly adapted from [17]. Section 5 summarized algorithmic procedures for finding optimal decisions. Conclusions are made in Section 6.

2 Notations and Preliminaries

In this note we consider Markov decision processes with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$ in continuous-time. In particular, the development of the considered Markov decision process $X = \{X(t), t \geq 0\}$ (with finite state space $\mathcal{I}$) over time is governed by the transition rates $q(j|i, a)$, for $i, j \in \mathcal{I}$, depending on the selected action $a \in \mathcal{A}_i$. For $j \neq i$ $q(j|i, a)$ is the transition rate from state i into state j, $q(i|i, a) = \sum_{j \in \mathcal{I}, j \neq i} q(j|i, a)$ is the transition rate out of state i. As concerns reward rates, $r(i)$ denotes the rate earned in state $i \in \mathcal{I}$, and $r(i, j)$ is the transition reward accrued to a transition from state i to state j.

Let $\xi(t) := \int_0^t r(X(\tau)\mathrm{d}\tau + \sum_{k=0}^{N(t)} r(X(\tau^-), X(\tau^+))$, obviously $\xi_(t)$ is the (random) reward obtained up to time t, where $X(t)$ denotes the state at time t, $X(\tau^-)$, $X(\tau^+)$ is the state just prior and after the kth jump, and $N(t)$ is the number of jumps up to time t. Similarly $\xi(t', t) := \int_{t'}^t r(X(\tau))\mathrm{d}\tau + \sum_{k=N(t')}^{N(t)} r(X(\tau^-), X(\tau^+))$ is the

[1]Institute of Information Theory and Automation of the Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic, sladky@utia.cas.cz

total (random) reward obtained in the time interval $[t', t)$; hence $\xi_(t+\Delta) = \xi(\Delta) + \xi(\Delta, t+\Delta)$ or $\xi(t+\Delta) = \xi(t) + \xi(t, t+\Delta)$.

In this note we shall suppose that the obtained random reward, say ξ, is evaluated by an exponential utility function, say $u^\gamma(\cdot)$, i.e. utility functions with constant risk sensitivity depending on the value of the risk sensitivity coefficient γ.

In case that $\gamma > 0$ (the risk seeking case) the utility assigned to the (random) reward ξ is given by $u^\gamma(\xi) := \exp(\gamma\xi)$, if $\gamma < 0$ (the risk averse case) the utility assigned to the (random) reward ξ is given by $u^\gamma(\xi) := -\exp(\gamma\xi)$, for $\gamma = 0$ it holds $u^\gamma(\xi) = \xi$ (risk neutral case). Hence we can write

$$u^\gamma(\xi) = \operatorname{sign}(\gamma)\exp(\gamma\xi) \tag{1}$$

and for the expected utility we have (E is reserved for expectation)

$$\bar{U}^{(\gamma)}(\xi) := \mathsf{E}u^\gamma(\xi) = \operatorname{sign}(\gamma)\mathsf{E}[\exp(\gamma\xi)], \quad \text{where} \quad \mathsf{E}\left[\exp(\gamma\xi)\right] = \sum_{k=0}^{\infty} \mathsf{E}\,\frac{1}{k!}(\gamma\xi)^k. \tag{2}$$

Then for the corresponding certainty equivalent $Z^\gamma(\xi)$ we have

$$u^\gamma(Z^\gamma(\xi)) = \operatorname{sign}(\gamma)\mathsf{E}[\exp(\gamma\xi)] \Longleftrightarrow Z^\gamma(\xi) = \gamma^{-1}\,\ln\{\mathsf{E}\left[\exp(\gamma\xi)\right]\}. \tag{3}$$

From (2),(3) we immediately conclude that

$$Z^\gamma(\xi) \approx \mathsf{E}\xi + \frac{\gamma}{2}\mathsf{Var}\,\xi. \tag{4}$$

In this note we focus attention on properties of the expected utility and the corresponding certainty equivalents if the stream of rewards generated by continuous-time Markov reward chain is evaluated by exponential utility functions. Recall that exponential utility function is separable, i.e. $u^\gamma(\xi^{(1)} + \xi^{(2)}) = \operatorname{sign}(\gamma)u^\gamma(\xi^{(1)}) \cdot u^\gamma(\xi^{(1)})$, what is very important for sequential decision problems. For what follows we shall need some more notions and notation.

A (Markovian) policy controlling the decision process is given as a piecewise constant right continuous function of time. In particular, $\pi = f(t)$, is a piecewise constant, right continuous vector function where $f(t) \in \mathcal{F} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$, and $f_i(t) \in \mathcal{A}_i$ is the decision (or action) taken at time t if the process $X(t)$ is in state i. Since π is piecewise constant, for each π we can identify the time points $0 < t_1 < t_2 \ldots < t_i < \ldots$ at which the policy switches; we denote by $f^i \in \mathcal{F}$ the decision rule taken in the time interval $(t_{i-1}, t_i]$. Policy which takes at all times the same decision rule, i.e. $\pi \sim f$, is called stationary; $Q(f)$ is the transition rate matrix with elements $q(j|i, f_i)$.

Let for $f \in \mathcal{F}$ $Q(f) = [q_{ij}(f_i)]$ be an $N \times N$ matrix whose ijth element $q_{ij}(f_i) = q(j|i, f_i)$ for $i \neq j$ and for the iith element we set $q_{ii}(f_i) = -q(i|i, f_i)$. The sojourn time of the considered process X in state $i \in \mathcal{I}$ is exponentially distributed with parameter $[q(i|i, f_i)]$. Hence the expected value of the reward obtained in state $i \in \mathcal{I}$ equals $\tilde{r}_i(f_i) = [q(i|i, f_i)]^{-1}\, r(i) + \sum_{j\in\mathcal{I}, j\neq i} q(j|i, f_i)\, r(i, j)$ and $\tilde{r}(f)$ is the (column) vector of reward rates at time t.

Recall that the row sums of a transition rate matrix $Q(f)$ are equal to null. Hence $\rho(f) = 0$ is an eigenvalue of $Q(f)$, the corresponding eigenvector is a unit vector. Moreover, the real part of every other eigenvalue of $Q(f)$ is negative. In particular, if for some $i \in \mathcal{I}$ and $f \in \mathcal{F}$ it happens that $\sum_{j\neq i} q_{ij}(f_i) < q_{ii}(f_i)$ on reaching state i the process X stops with probability $q_{ii}(f_i) = -\sum_{j\neq i} q_{ij}(f_i)$ and if $Q(f)$ is irreducible then for the spectral radius of $\tilde{Q}(f) = Q(f)$ it holds $\tilde{\rho}(f) < 0$ and the real part of every other eigenvalue of $Q(f)$ is less than $\tilde{\rho}(f)$. Observe that $[I - \tilde{Q}(f)]^{-1} = \sum_{k=0}^{\infty}[\tilde{Q}(f)]^k$ exists[2]. It can be shown (see e.g. Gantmakher [1]) that the matrix $\tilde{Q}(f)$ is nonsingular.

Using policy $\pi = f(t)$ means that if the Markov chain X was found to be in state i at time t, the action chosen at this time is $f_i(t)$, i.e. the ith coordinate of $f(t) \in \mathcal{F}$. For any policy $\pi = f(t)$ the accompanying set of transition rate matrices $\{Q(f(t)), t \geq 0\}$ determines a continuous-time (in general, nonstationary) Markov process.

Let $P(\cdot, \cdot, \pi)$ be the $N \times N$ matrix of transition functions associated with Markov chain X, i.e. for each $0 \leq s \leq t$ the ijth element of $P(\cdot, \cdot, \pi)$, denoted $P_{ij}(s, t, \pi)$, is the probability that the chain is in state j at time t given it was in state i at time s and policy π is followed. Obviously, by the well-known Chapman–Kolmogorov equation $P(s, t, \pi) = \sum_{u\in\mathcal{I}} P(s, u, \pi)\, P(u, t, \pi)$ for each $0 \leq s < u < t$. The values $P(s, t, \pi)$ are absolutely continuous in t and satisfy the system of differential equations (except possibly where the piecewise constant policy switches)

$$\frac{\partial P(s, t, \pi)}{\partial t} = P(s, t, \pi)\, Q(f(t)), \quad \frac{\partial P(s, t, \pi)}{\partial s} = -Q(f(s))P(s, t, \pi) \tag{5}$$

[2]In this note the symbol I is reserved for unit matrix, similarly e is reserved for a unit column vector.

where $P(s,s,\pi) = I$. In what follows it will be often convenient to let $P(t,\pi) = P(0,t,\pi)$. By (5) we then immediately get for any $t \geq 0$

$$\frac{\mathrm{d}P(t,\pi)}{\mathrm{d}t} = P(t,\pi)\,Q(f(t)) \quad \text{along with} \quad P(t,\pi) = I + \int_0^t P(u,\pi)Q(f(u))\mathrm{d}u. \tag{6}$$

Moreover, if the considered time horizon t tends to null, i.e. if for any piecewise constant policy $\pi = f(t)$ the considered time horizon $\Delta \downarrow 0$, for the ij-th element of $P(\Delta,\pi)$ we have

$$P_{ij}(\Delta,\pi) = \begin{cases} 1 + q_{ii}(f_i(t))\Delta + o(\Delta^2) & \text{for } i = j \\ q_{ij}(f_i(t))\Delta + o(\Delta^2) & \text{for } i \neq j \end{cases} \tag{7}$$

It is well known that for any stationary policy $\pi \sim f$ there exists $\lim_{t\to\infty} P(t,\pi) = P^*(\pi)$ and, moreover, that for any $t \geq 0$ (for stationary policy $\pi \sim f$ we write f instead of π)

$$P(t,f)\,P^*(f) = P^*(f)\,P(t,f) = P^*(f)\,P^*(f) = P^*(f), \tag{8}$$

$$Q(f)\,P^*(f) = P^*(f)\,Q(f) = 0 \quad \text{where} \quad P(t,f) = \exp(Q(f)t) = \sum_{k=0}^{\infty} \frac{1}{k!}[Q(f).t]^k. \tag{9}$$

3 Exponential Utility and the Corresponding Moments

Supposing that stationary policy $\pi \sim f$ is followed, let $U_i^{(\gamma)}(t,f) := \mathsf{E}_i^f \mathrm{e}^{\gamma\xi(t)}$ and by (2) $\bar{U}_i^{(\gamma)}(t,f) := \mathsf{E}_i^f u^\gamma(\xi(t))$ be the expected value of the random reward evaluated according the exponential utility function $u^\gamma(\cdot)$ earned up to time t if the process starts in state i (obviously $\bar{U}_i^{(\gamma)}(t,f) = \mathrm{sign}(\gamma)\,U_i^{(\gamma)}(t,f)$). Recall that $\mathsf{E}_i^f[\exp(\gamma\xi(t))] = U_i^{(\gamma)}(t,f)$ is also the moment generating function of $\xi(t)$. Hence (cf. (2)) for the k-th moment of $\xi(t)$ it holds

$$M_i^{(k)}(\xi(t),f) := \frac{\mathrm{d}^k}{\mathrm{d}\gamma^k}\mathsf{E}_i^f[\exp(\gamma\xi(t)))]|_{\gamma=0} = \mathsf{E}_i^f\,\xi(t))^{k-1} \tag{10}$$

and the Taylor expansion around $\gamma = 0$ reads

$$U_i^{(\gamma)}(\xi(t),f) = 1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} M_i^{(k)}(\xi(t),f) \;\text{ for }\; |\gamma| < h. \tag{11}$$

Since $u^\gamma(\cdot)$ is separable and multiplicative we have

$$u^\gamma(\xi(t)+\Delta) = \mathrm{sign}(\gamma)\,u^\gamma(\xi(\Delta))\cdot u^\gamma(\xi(\Delta,t)), \quad \text{or} \quad u^\gamma(\xi(t+\Delta)) = \mathrm{sign}(\gamma)\,u^\gamma(\xi(t))\cdot u^\gamma(\xi(t,t+\Delta)).$$

On taking expectations we immediately conclude that for $U_i^{(\gamma)}(t,f) := \mathsf{E}_i^f\{\mathrm{e}^{\gamma\xi(t)}\}$ we have (δ_{ij} is the Kronecker symbol)

$$U_i^{(\gamma)}(t+\Delta,f) = \sum_{j=1}^{N} P_{ij}(\Delta,f)\cdot[\mathrm{e}^{\gamma r(i)\Delta}\delta_{ij} + \mathrm{e}^{\gamma r(ij)}(1-\delta_{ij})]\cdot U_j^{(\gamma)}(t,f). \tag{12}$$

Since

$$\mathrm{e}^{\gamma r(i)\Delta} = 1 + \gamma r(i)\Delta + o(\Delta^2), \quad P_{ij}(\Delta,f) = \begin{cases} 1 + q_{ii}(f_i)\Delta + o(\Delta^2) & \text{for } i = j \\ q_{ij}(f_i)\Delta + o(\Delta^2) & \text{for } i \neq j \end{cases}$$

on letting $\Delta \to 0+$ we conclude that

$$U_i^{(\gamma)}(t+\Delta,f) = (1+q_{ii}(f_i)\Delta)\mathrm{e}^{\gamma r(i)\Delta}U_i^{(\gamma)}(t,f) + \sum_{j=1,\,j\neq i}^{N} q_{ij}\Delta\mathrm{e}^{\gamma r(ij)}\cdot U_j^{(\gamma)}(t,f) + o(\Delta^2). \tag{13}$$

Since the process X is time homogeneous and $(1+\gamma r(i)\Delta)(1+q_{ii}(f_i)\Delta) = 1 + (q_{ii}(f_i) + \gamma r(i))\Delta + o(\Delta^2)$ after some manipulation we arrive at

$$\frac{\mathrm{d}U_i^{(\gamma)}(t,f)}{\mathrm{d}t} = (q_{ii}(f_i) + \gamma r(i))U_i^{(\gamma)}(t,f) + \sum_{j=1,j\neq i}^{N} q_{ij}(f_i)\mathrm{e}^{\gamma r(ij)}\cdot U_j^{(\gamma)}(t,f) \tag{14}$$

that can be also written in matrix form as

$$\frac{\mathrm{d}U^{(\gamma)}(t,f)}{\mathrm{d}t} = \bar{Q}^{(\gamma)}(f) \cdot U^{(\gamma)}(t,f) \tag{15}$$

where $U^{(\gamma)}(t,f) = [U_i^{(\gamma)}(t,f)]$ is a (column) vector, and $\bar{Q}^{(\gamma)}(f) = [\bar{q}_{ij}^{(\gamma)}(f_i)]$ is an $N \times N$ matrix with nonnegative off-diagonal elements

$$\bar{q}_{ij}^{(\gamma)}(f_i) = \begin{cases} q_{ii}(f_i) + \gamma \cdot r(i) & \text{for } i = j \\ q_{ij}(f_i) \cdot \mathrm{e}^{\gamma\, r(i,j)} & \text{for } i \neq j \end{cases}$$

Hence by (15) if $U^{(\gamma)}(0,f) = e$

$$U^{(\gamma)}(t,f) = \exp[\bar{Q}^{(\gamma)}(f).t] \cdot e = \sum_{k=0}^{\infty} \frac{1}{k!} [\bar{Q}^{(\gamma)}(f).t]^k \cdot e. \tag{16}$$

To study asymptotic behavior of $U^{(\gamma)}(t,f)$ first recall (cf. Gantmakher [1]) that for any matrix with non-negative off-diagonal elements there exists a positive eigenvalue (called the dominant eigenvalue) such that the real part of any other eigenvalue is nongreater than the dominant eigenvalue. In addition, the corresponding eigenvector can be selected positive and even strictly positive if the considered matrix is irreducible or at least unichain. In particular, let $\sigma^{(\gamma)}(f)$ be the dominant eigenvalue of $\bar{Q}^{(\gamma)}(f)$ and $v^{(\gamma)}(f)$ the corresponding eigenvector, i.e.

$$\bar{Q}^{(\gamma)}(f) \cdot v^{(\gamma)}(f) = \sigma^{(\gamma)}(f) \cdot v^{(\gamma)}(f).$$

Observe that if the matrix $Q(f)$ is unichain then $\bar{Q}^{(\gamma)}(f)$ remains unichain at least if risk-sensitive coefficient γ is sufficiently close to null.

4 Mean Variance Optimality

Another approach how to handle variability-risk features in continuous-time Markov decision processes is based on mean variance selection rules, i.e. we optimize the weighted sum of average or total reward and its variance. To this end, we focus attention on undiscounted continuous-time Markov decision chains, i.e. we assume that the risk sensitive coefficient $\gamma = 0$, hence the corresponding utility $u^{(0)}(\xi) = \xi$ and the expected utility $\bar{U}^{(0)}(\xi) := \mathsf{E}\xi$. Hence for the selected long-run stationary policy the variability is measured either by the ratio of long-run average reward and long run average variance. Of course, it is optimal in the class of policies with a fixed average reward to select policy with minimal average variance. This method that transforms finding minimal variance to finding minimal average reward of suitably transformed Markov decision process was initially suggested in [7] for discrete-time Markov decision chains, similar procedure for continuous-time models were used in [3],[4],[17].

To begin with, we start with well-known approach for evaluating and finding optimal average policy for continuous-time Markov decision processes (see e.g. [5, 2]). Similar approach can be also used for finding second moment and the corresponding variance of the considered Markov reward chain.

Considering the risk-neutral models, on taking expectations we conclude that for $U_i^{(0)}(t,f) := \mathsf{E}_i^f\{\xi(t)\}$ we have

$$U_i^{(0)}(t+\Delta,f) = r(i) \cdot \Delta + (1 + q_{ii}(f_i) \cdot \Delta) U_i^{(0)}(t,f) + \Delta \sum_{j=1, j\neq i}^{N} q_{ij}(f_i)[r(ij) + U_j^{(0)}(t,f)] + o(\Delta^2) \tag{17}$$

Since the process X is time homogeneous, after some manipulations, on letting $\Delta \to 0+$, we arrive at

$$\frac{\mathrm{d}U_i^{(0)}(t,f)}{\mathrm{d}t} = r(i) + \sum_{j=1, j\neq i}^{N} q_{ij}(f_i) \cdot r(ij) + \sum_{j=1, j\neq i}^{N} q_{ij}(f_i) \cdot [U_j^{(0)}(t,f) - U_i^{(0)}(t,f)] \tag{18}$$

It is well-known from the literature if the considered continuous-time Markov chain has a single class of recurrent states that the long-run average reward is independent on the initial state and $g(f) = \lim_{t\to\infty} \frac{1}{t} U_i^{(0)}(t,f)$ exists and is independent of the starting state. In addition, it is well-known that there exist vector $w(f)$ (with elements $w_i(f)$) such that

$$U_i^{(0)}(t,f) = g(f) \cdot t + w_i(f) - [P(t,f)w(f)]_i \qquad ([w]_i \text{ denotes the } i\text{th entry of the column vector } w).$$

In what follows we shall be also interested in the second moment and the corresponding variance of random reward $\xi(t)$. Since $\mathsf{E}_i^f[\xi(t+\Delta)]^2 = \mathsf{E}_i^f[\xi(t)]^2 + 2\cdot\mathsf{E}_i^f\xi(t)\cdot\Delta + \mathsf{E}_i^f[\Delta]^2$, for the second moment of $\xi(t)$, say $S_i(f)(\xi) := \mathsf{E}_i^f(\xi)^2$, we get

$$\begin{aligned} S_i(t+\Delta,f) &= (1+\Delta\cdot q_{ii}(f_i))\left\{2\Delta\cdot r(i)U_i^{(0,f)}(t,f)+S_i(t,f)\right\} \\ &\quad +\Delta\sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\left\{[r(ij)]^2+2r(ij)U_j^{(0)}(t,f)+S_j(t,f)\right\}+o(\Delta^2). \end{aligned} \tag{19}$$

Hence (similarly to (18)) for $\Delta\to 0+$ from (19) we get

$$\begin{aligned} \frac{\mathrm{d}S_i(t,f)}{\mathrm{d}t} &= 2r(i)U_i^{(0)}(t,f)+\sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\left\{[r(ij)]^2+2r(ij)U_j^{(0)}(t,f)\right\} \\ &\quad +\sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)[S_j(t,f)-S_i(t,f)]. \end{aligned} \tag{20}$$

Since $\sigma_i(t,f) = S_i(t,f) - [U_i^{(0)}(t,f)]^2$ we thus obtain:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\sigma_i(t,f) &= \frac{\mathrm{d}}{\mathrm{d}t}S_i(t,f) - 2U_i^{(0)}(t,f)\frac{\mathrm{d}}{\mathrm{d}t}U_i^{(0)}(t,f) \\ &= 2r(i)U_i^{(0)}(t,f)+\sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\left\{[r(ij)^2+2r(ij)U_j^{(0)}(t,f)\right\}+\sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)[S_j(t,f)-S_i(t,f)] \\ &\quad -2U_i^{(0)}(t,f)\left\{r(i)+\sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)r(ij)+\sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)[U_j^{(0)}(t,f)-U_i^{(0)}(t,f)]\right\} \end{aligned} \tag{21}$$

By substituting $S_j(t,f) = \sigma_j(t,f) + [U_j^{(0)}(t,f)]^2$ (21) can be rewritten as:

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma_i(t,f) = \sum_{j\in\mathcal{I}} q_{ij}(f_i)\sigma_j(t,f) - 2U_i^{(0)}(t,f)\left\{\sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)r(ij)+\sum_{j\in\mathcal{I}} q_{ij}(f_i)U_j^{(0)}(t,f)\right\} \tag{22}$$

Using a more detailed analysis (see [17]) it can be shown that

$$G(f) := \lim_{t\to\infty}\frac{1}{t}\sigma_i(t,f) \tag{23}$$

exists and is independent of the initial state i. For unichain models the values $G(f)$ of the average variance can be calculated as average reward of a continuous time Markov chain where one stage reward in state i, say $s_i(f)$, is equal to

$$s_i(f) = \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\{[r(ij)+w_j(f)]^2-[w_i(f)]^2\}+2[r(i)-g(f)]w_i(f), \text{ or} \tag{24}$$

$$s_i^{(1)}(f) = \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\{[r(ij)+w_j(f)]^2-[w_i(f)]^2\}+2r(i)w_i(f), \text{ or} \tag{25}$$

$$s_i^{(2)}(f) = \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\{[r(ij)]^2+2r(ij)\}+2r(i)w_i(f). \tag{26}$$

5 Finding optimal control policy

Considering continuous-time Markov decision processes, it has been shown in section 3 that the growth of risk-sensitive optimality is governed by matrices with nonnegative off-diagonal elements. Recalling that for any matrix with nonnegative off-diagonal elements there exists a real eigenvalue (called the dominant eigenvalue) such that the real part of any eigenvalue of the matrix is nongreater than the dominant eigenvalue, for maximizing the risk-sensitive average reward it suffices to find stationary policy defining feasible matrix with maximum possible real eigenvector. To this end we can employ policy iteration or value iteration specified in [14]. This methods also enable to generated upper and lower estimates on the maximum possible dominant eigenvalue in each iteration step.

Another approach for finding optimal policies was studied in section 4. Using this approach the decision maker selects by standard policy or value iteration methods the set of all (non-randomized) stationary policies with maximal average reward. In the next step in the class of stationary policies maximizing the average reward the decision maker selects the policy with minimal average variance. Observe that the construction of the policy with minimal average variance is limited only on the class of optimal policies (resp. another selected class of policies) since the formulas for construction the policy with minimal variance heavily employ values of some coefficients constructed for optimal policies (cf. formulas (24)–(26)). Up to now we looked for policies with minimal variance in the class of policies with maximal total reward. However, considering stationary policy (even randomized), not maximizing the average reward, policy and value iteration methods can be used for finding policies guaranteeing the desired reward with minimal possible variance.

6 Conclusions

In this note we studied risk-sensitive average optimality in risk-sensitive continuous-time Markov decision chains. The obtained results extend some older result reported for uncontrolled models in [13] and are overlapping with recent results reported in [18] obtained by different methods and are similar to the results reported in [3] and [4].

The reported result are also confronted with so-called mean-variance optimality criterion for finding policies with minimal average variance in certain class of policies (usually in the class of policies with maximal average reward). The results reported in section 4 are similar to Thm.10.5 in [2], Thm.3 in [3] or Thm.3.4 in [10] (in these references no transient rewards are considered).

Acknowledgement

This research was supported by the Czech Science Foundation under Grant 18-02739S.

References

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

[2] Guo, X. and Hernández-Lerma, O. (2009). *Continuous-Time Markov Decision Processes: Theory and Applications.* Berlin: Springer.

[3] Guo, X. and Song, X. (2009). Mean-variance criteria for finite continous-time Markov decision processes. *IEEE Trans. Automat. Control* 54, 2151–2157.

[4] Guo, X., Ye, L. and Yin, G.A. (2012). Mean-variance optimization problem for discounted Markov decision processes. *European J. Oper. Res.* 220, 423–429.

[5] Howard, R.A. (1960). *Dynamic Programing and Markov Processes.* New York: Wiley.

[6] Howard, R.A. and Matheson, J. (1972). Risk-sensitive Markov decision processes. *Manag. Science* 16, 356–369.

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

[8] Markowitz, H. (1952). Portfolio Selection. *Journal of Finance* 7, 77–92.

[9] Markowitz, H. (1959). *Portfolio Selection - Efficient Diversification of Investments.* New York: Wiley.

[10] Prieto-Rumeau, T. and Hernández-Lerma, O. (2009). Variance minimization and the overtaking optimality approach to continuous-time controlled Markov chains. *Math. Method Oper. Res.* 70, 527–540.

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

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

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

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

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

[16] Sladký, K. (2017). Second order optimality in Markov decision chains. *Kybernetika* 53, 1086–1099.

[17] Van Dijk, N.M. and Sladký, K. (2006). On total reward variance for continuous-time Markov reward chains. *J. Appl. Probab.* 43, 1044–1052.

[18] Wei, Q. and Chen, X. (2016). Continuous-time Markov decision processes under risk-sensitive average cost criterion. *Oper. Res. Lett.* 44, 457–462.