Central Moments and Risk-Sensitive Optimality in Markov Reward Processes

Karel Sladký [1]

Abstract. In this note we consider discrete- and continuous-time Markov decision processes with finite state spaces. There is no doubt that usual optimality criteria examined in the literature on optimization of Markov reward processes, e.g. total discounted or mean reward, may be quite insufficient to select more sophisticated criteria that reflect also the variability-risk features of the problem. In this note we focus on models 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). For the risk-sensitive case, i.e. if the considered risk-sensitivity coefficient is non-zero, we establish explicit formulas for growth rate of expectation of the exponential utility function. Recall that in this case along with the total reward also its higher moments are taken into account. Using Taylor expansion of the utility function we present explicit formulae for calculating variance and higher central moments of the total reward generated by the Markov reward process along with its asymptotic behavior.

Keywords: Discrete- and continuous-time Markov reward chains, exponential utility, moment generating functions, formulae for central moments.

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 robustness of 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 stability and variability-risk features of the problem. Perhaps the best known approaches stem from the classical work of Markowitz (1952) on mean variance selection rules, i.e. we optimize the weighted sum of average or total reward and its variance, and from the seminal paper titled "Risk-sensitive Markov decision processes" of Howard and Matheson (1972), based on evaluating generated reward by exponential utility functions. Higher moments and variance of cumulative rewards in Markov reward chains have been originally studied for discrete time models. Research in this direction has been initiated in early papers Mandl (1971), Jaquette (1973) and Sobel (1982). For connections with risk sensitive models see e.g. Cavazos-Cadena and Fernandez-Gaucherand (1999), Cavazos-Cadena and Hernández-Hernández (2005) and Sladký (2005),(2008),(2018) and (2020). For basic facts on controlled Markov processes, see e.g. Puterman (1994) or Ross (1983).

The present paper is structured as follows. Section 2 contains notations and summary of basic facts on discrete- and continuous-time Markov reward processes. Discrete-time Markov models with exponential utility function (called risk-sensitive Markov chains) are studied in section 3 along the corresponding moment generating functions and explicit formulas of higher moments and higher central moments. Similar results for continuous-time Markov reward chains are sketched in Section 4.

2 Notations and Preliminaries

In this note we consider Markov decision processes with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$ evolving in discrete- and continuous-time.

In the discrete-time case, we consider Markov decision chain $X^{\mathrm{d}} = \{X_n, n = 0, 1, \ldots\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$, and finite set $\mathcal{A}_i = \{1, 2, \ldots, K_i\}$ of possible decisions (actions) in state $i \in \mathcal{I}$. Supposing that in state $i \in \mathcal{I}$ action $a \in \mathcal{A}_i$ is selected, then state j is reached in the next transition with a given probability $p_{ij}(a)$ and one-stage transition reward r_{ij} will be accrued to such transition. Obviously, $r_i(a) = \sum_{j \in \mathcal{I}} p_{ij}(a) r_{ij}$ is the expected one-stage reward.

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

In the continuous-time setting, the development of the considered Markov decision process $X^{\rm c} = \{X(t), t \geq 0\}$ (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, $q(i|i,a) = -\sum_{j\in\mathcal{I}, j\neq i} q(j|i,a)$. Recall that on entering state i the process stays in state i for a random time that is exponentially distributed with parameter $q(i,a) = -q(i|i,a)$ and the next jump to state j occurs with probability $p_{ij}(a) = q(j|i,a)/q(i,a)$. As concerns reward rates, $r(i)$ denotes the rate earned in state $i \in \mathcal{I}$, and $r(i,j)$ is the transition rate accrued to a transition from state i to state j.

A (Markovian) policy controlling the decision process is given either by a sequence of decision at every time point (discrete-time case) or as a piecewise constant right continuous function of time (continuous-time case). Policy which takes at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary; in discrete-time models $P(f)$ is the transition probability matrix with elements $p_{ij}(f_i)$. Obviously, $r_i(f_i) = \sum_{j=1}^{N} p_{ij}(f_i) r_{ij}$ is the expected one-stage reward obtained in state $i \in \mathcal{I}$ and $r(f)$ denotes the corresponding N-dimensional column vector of one-stage rewards. $v(f,n) := \sum_{i=0}^{n-1} [P(f)]^n \cdot r(f)$ is the (column) vector of total rewards accrued after $n+1$ transitions, its ith entry $v_i(f,n)$ denotes expectation of the total reward if the process $X^{\rm d}$ starts in state i. Then $g(f) = P^*(f) \cdot r(f)$ (where $P^*(f) = \lim_{m\to\infty} m^{-1} \sum_{n=0}^{m-1} P^n(f)$) is the (column) vector of average rewards, its ith entry $g_i(f)$ denotes the average reward if the process starts in state i. If the Markov chain $P(f)$ is unichain, i.e. it has a single class of recurrent states, then $g(f)$ is constant vector of average rewards with elements $\bar{g}(f)$.

Similarly, for the continuous-time case policy controlling the chain, $\pi = f(t)$, 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. $Q(f) = [q_{ij}(f_i)]$ for $f \in \mathcal{F}$ is 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 $q_{ii}(f_i) = -q(i|i,f_i)$ (recall that the row sums in a transition rate matrix $Q(f)$ are equal to null). Expected value of the reward obtained in state $i \in \mathcal{I}$ equals $r_i(f_i) = [q(i|i,f_i)]^{-1} r(i) + \sum_{j\in\mathcal{I}, j\neq i} q(j|i,f_i)\, r(i,j)$ and $r(f) = [r_i(f)]$ is the (column) vector of reward rates at time t.

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

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

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

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

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

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

From (2),(3) we immediately conclude

$$Z^\gamma(\xi) \approx \mathsf{E}\xi + \frac{\gamma}{2} \mathsf{Var}\, \xi \quad \text{where} \quad \mathsf{Var}\, \xi = \mathsf{E}[\xi - \mathsf{E}\xi]^2. \tag{4}$$

In particular, considering discrete-time models, let ξ_n be the random reward obtained in the first n transitions of the considered Markov process, i.e.

$$\xi_n = r_{X_0,X_1} + r_{X_1,X_2} + \ldots + r_{X_{n-1},X_n}. \tag{5}$$

Supposing that stationary policy $\pi \sim (f)$ is followed and the process starts in state $i = X_0$ then $\mathsf{E}_i^\pi \xi_n = v_i(f,n)$ where $v(f,n) = [P(f)]^n r(f)$.

Similarly, for the continuous-time models, the (random) reward obtained in the first n jumps of the process depends not only on transition rewards (say r_{X_1,X_2}) connected with jumps of the process, but also on reward rates (say $r_{(\cdot)}$) and random times (say $\tau_\cdot$) spent in the states, and is equal to

$$\xi_n = r_{X_0} \cdot \tau_{X_0} + r_{X_0,X_1} + r_{X_1} \cdot \tau_{X_1} + r_{X_1,X_2} + \ldots + r_{X_{n-1}} \cdot \tau_{X_{n-1}} + r_{X_{n-1},X_n}.$$

3 Discrete-Time Models: Exponential Utility and Higher Moments

It is well-known (see e.g. Puterman (1994), Ross (1983)) that if $P(f)$ is unichain then there exist vector $w(f)$ (unique up to constant vector) and constant vector $g(f)$ (with elements $\bar{g}(f)$) such that $w(f) + g(f) = r(f) + P(f)w(f)$. Then we can conclude that the vector of total expected rewards

$$v(f,n) = w(f) + n \cdot g(f) - [P(f)]^n \cdot w(f), \text{ in particular, } v_i(f,n) = w_i(f) + n \cdot \bar{g}(f) - [[P(f)]^n \cdot w(f)]_i, \quad (6)$$

i.e. the growth rate of $v_i(f,n)$ is linear over n.

Moreover, if the process starts in state i and stationary policy $\pi \sim (f)$ is followed from (2), (5) we can conclude that the growth of expected utility for risk-sensitive models in the n steps

$$U_i^{(\pi,\gamma)}(n) := \mathsf{E}_i^{(\pi)} \, \mathrm{e}^{\gamma(r_{X_0,X_1} + r_{X_1,X_2} + \ldots + r_{X_{n-1},X_n})} \quad (7)$$

well corresponds to the growth of $v_i(f,n)$. Observe that for γ near to zero $U_i^{(\pi,\gamma)}(\xi_n)$ risk-sensitive models can be well approximated by $v_i(f,n)$ of the classical models.

In what follows we show that if transition probability matrix $P(f) = [p_{ij}(f)]$ is unichain then:

There exist real $g^\gamma(f)$, $w_i^\gamma(f)$'s ($i \in \mathcal{I}$) such that for $\tilde{\varphi}_{ij}(w,g,f) := r_{ij} - g^\gamma(f) + w_j^\gamma(f) - w_i^\gamma(f)$

$$\sum_{j \in \mathcal{I}} p_{ij}(f_i) \, \mathrm{e}^{\gamma[r_{ij} - w_i^\gamma(f) + w_j^\gamma(f) - g^\gamma(f)]} = 1 \Longleftrightarrow \sum_{j \in \mathcal{I}} p_{ij}(f_i) \, \mathrm{e}^{\gamma[r_{ij} + w_j^\gamma(f)]} = \mathrm{e}^{\gamma[g^\gamma(f) + w_i^\gamma(f)]} \quad \text{for all } i \in \mathcal{I} \quad (8)$$

To verify (9), let $m_{ij}(f) := p_{ij}(f) \cdot \mathrm{e}^{\gamma r_{ij}}$, $\rho(f) := \mathrm{e}^{\gamma \cdot g^\gamma(f)}$, $x_j(f) := \mathrm{e}^{\gamma \cdot w_j^\gamma(f)}$.

Then by the Perron–Frobenius theorem for nonnegative matrices (see e.g. Gantmacher (1959))

$$\sum_{i \in \mathcal{I}} m_{ij}(f) \cdot x_j(f) = \rho(f) \cdot x_i(f), \qquad i \in \mathcal{I} \quad (9)$$

where $M(f) = [m_{ij}(f)]_{i,j}$ is an irreducible nonnegative matrix (or reducible nonnegative matrix with strictly positive right Perron eigenvector), $\rho(f)$ is the eigenvalue of $M(f)$ (equal to the spectral radius of $M(f)$) and $[x_i(f)]_i$ is the corresponding right eigenvector of $M(f)$.

Similarly, from (7) we conclude that for $\pi \sim (f)$

$$U_i^{(\pi,\gamma)}(n) = \mathsf{E}_i^{\pi} \, \mathrm{e}^{\gamma \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}} = \mathrm{e}^{\gamma[n g^\gamma(f) + w_i^\gamma(f)]} \times \mathsf{E}_i^{\pi} \, \mathrm{e}^{\gamma[\sum_{k=0}^{n-1} \tilde{\varphi}_{X_k,X_{k+1}}(w^\gamma(f),\gamma(f)) - w_{X_n}^\gamma(f)]}. \quad (10)$$

Observe that the first term on the RHS of (10) is non-random and hence (for $|c| \le w_i(f)$)

$$\mathsf{E}_i^{\pi} \, \mathrm{e}^{\gamma[\sum_{k=0}^{n-1} \tilde{\varphi}_{X_k,X_{k+1}}(w^\gamma(f),g^\gamma(f)) - c]} \le \frac{U_i^{(\pi,\gamma)}(n)}{\mathrm{e}^{\gamma[n g^\gamma(f) + w_i^\gamma(f)]}} \le \mathsf{E}_i^{\pi} \, \mathrm{e}^{\gamma[\sum_{k=0}^{n-1} \tilde{\varphi}_{X_k,X_{k+1}}(w^\gamma(f),g^\gamma(f)) + c]} \quad (11)$$

If (stationary) policy $\pi \sim (f)$ is followed then by (2) $U_i^\pi(\gamma,n) = \mathsf{E}_i^{\pi} \, [\exp(\gamma \xi^{(n)})]$ is also the moment generating function of $\xi^{(n)}$. Hence (cf. (2)) for some $h > 0$ and any $|\gamma| < h$

$\frac{\mathrm{d}}{\mathrm{d}\gamma} \mathsf{E}_i^{\pi} \, [\exp(\gamma \xi^{(n)})] = \mathsf{E}_i^{\pi} \, \xi^{(n)} [\exp(\gamma \xi^{(n)})]$, hence for $k = 0,1,2\ldots$, $n = 0,1,2\ldots$

$$M_i^{(\pi,k)}(n) := \mathsf{E}_i^{\pi} \left(\xi^{(n)}\right)^k = \frac{\mathrm{d}^k}{\mathrm{d}\gamma^k} \mathsf{E}_i^{\pi} \, [\exp(\gamma \xi^{(n)})]|_{\gamma=0} \text{ is the } k\text{th moment of } \xi^{(n)} \quad (12)$$

and (cf. (2)) the Taylor expansion around $\gamma = 0$ reads for $|\gamma| < h$

$$U_i^\pi(\gamma,n) = 1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} M_i^{(\pi,k)}(n) \text{ and } \mathrm{e}^{\gamma r_{ij}} = 1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} [r_{ij}]^k. \quad (13)$$

From (12),(13) we immediately get

$$1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} M_i^{(\pi,k)}(n+1) = \sum_{j \in \mathcal{I}} p_{ij}(f) \left(1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} [r_{ij}]^k\right) \left(1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} M_j^{(\pi,k)}(n)\right). \quad (14)$$

Similarly on introducing the moment generating function for the central moments of $\xi^{(n)}$ by

$$\tilde{U}_i^{\pi}(\gamma, n) := \mathsf{E}_i^{\pi}\,[\exp(\gamma(\xi^{(n)} - \mathsf{E}_i^{\pi}\,\xi^{(n)})] \quad \text{for all } \; i \in \mathcal{I} \tag{15}$$

for the kth central moment of $\xi^{(n)}$ we have

$$\widetilde{M}_i^{(k,\pi)}(n) := \mathsf{E}_i^{\pi}\,\left[\xi^{(n)} - \mathsf{E}_i^{\pi}\,\xi^{(n)}\right]^k = \frac{\mathrm{d}^k}{\mathrm{d}\gamma^k}\mathsf{E}_i^{\pi}\,[\exp(\gamma(\xi^{(n)} - \mathsf{E}_i^{\pi}\,\xi^{(n)})]|_{\gamma=0} \tag{16}$$

and the Taylor expansion around $\gamma = 0$ for $|\gamma| < h$ reads

$$\tilde{U}_i^{\pi}(\gamma, n) = 1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} \cdot \widetilde{M}_i^{(\pi,k)}(n). \tag{17}$$

For the central moments similarly to (15), (16) we can conclude from (10), (11) that

$$\tilde{U}_i^{\pi}(\gamma, n) := \mathsf{E}_i^{\pi}\,\mathrm{e}^{\gamma[\xi^{(n)} - (ng - w_i + w_{X_n})]} = \sum_{j \in \mathcal{I}} p_{ij}(f) \cdot \mathrm{e}^{\gamma(r_{ij} - g + w_i - w_j)} \tilde{U}_j^{\pi}(\gamma, n-1) \tag{18}$$

where $\quad \tilde{U}_j^{\pi}(\gamma, n-1) = \mathsf{E}_j^{\pi} \mathrm{e}^{\gamma[\xi^{(1,n)} - (n-1)g + w_j - w_{X_n}]}$.

In analogy to (14) we get

$$1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} \widetilde{M}_i^{(\pi)}(k, n+1) = \sum_{j \in \mathcal{I}} p_{ij}(f) \big(1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} [r_{ij} - (g + w_i - w_j)]^k\big) \times \big(1 + \sum_{k=1}^{\infty} \frac{\gamma^k}{k!} \widetilde{M}_j^{(\pi)}(k, n)\big). \tag{19}$$

Similarly as in the previous section our analysis based on (15),(16),(17) and (18) enables to generate recursively all central moments of $\xi^{(n)}$.

By comparing in (18) the terms $\gamma^k \;\; (k = 1, 2, \ldots)$ we obtain the following recursive formulas for the central moments (obviously, the first central moment $\widetilde{M}_i^{(\pi,1)}(n) \equiv 0$ for all n). In particular,

$$\text{For } k = 2 : \widetilde{M}_i^{(\pi,2)}(n+1) = \sum_{j \in \mathcal{I}} p_{ij}(f)\,[(r_{ij} + w_j) - (g + w_i)]^2 + \sum_{j \in \mathcal{I}} p_{ij}(f)\,\widetilde{M}_j^{(\pi,2)}(n). \tag{20}$$

$$\begin{aligned}\text{For } k = 3 : \widetilde{M}_i^{(\pi,3)}(n+1) &= \sum_{j \in \mathcal{I}} p_{ij}(f)\,[(r_{ij} + w_j) - (g + w_i)]^3 \\ &+3 \sum_{j \in \mathcal{I}} p_{ij}(f)\,[(r_{ij} + w_j) - (g + w_i)]\,\widetilde{M}_j^{(\pi,2)}(n) + \sum_{j \in \mathcal{I}} p_{ij}(f)\,\widetilde{M}_j^{(\pi,3)}(n).\end{aligned} \tag{21}$$

In general:

$$\begin{aligned}\widetilde{M}_i^{(\pi,s)}(n+1) &= \sum_{j \in \mathcal{I}} p_{ij}(f)\;\{[(r_{ij} + w_j) - (g + w_i)]^s\} \\ &+ \sum_{j \in \mathcal{I}} p_{ij}(f)\;\left\{\sum_{k=1}^{s-1} \left(\begin{array}{c} s \\ k \end{array} \right)\,[(r_{ij} + w_j) - (g + w_i)]^k\,\widetilde{M}_j^{(\pi,s-k)}(n)\right\} + \sum_{j \in \mathcal{I}} p_{ij}(f)\,\widetilde{M}_j^{(\pi,s)}(n)\end{aligned} \tag{22}$$

that can be also written as

$$\widetilde{M}_i^{(\pi,s)}(n+1) = \sum_{k=0}^{s} \left(\begin{array}{c} s \\ k \end{array} \right) \sum_{j \in \mathcal{I}} p_{ij}(f)\left\{[(r_{ij} + w_j) - (g + w_i)]^k\,\widetilde{M}_j^{(\pi,s-k)}(n)\right\} \tag{23}$$

From these equations we immediately conclude that if stationary policy $\pi \sim (f)$ is followed the variance (i.e. the central second moment) of the total reward grows linearly over time and the growth rate $g^{(2)}$ of $\widetilde{M}_i^{(\pi,2)}(n)$ in (21) can be found as a solution of

$$\widetilde{M}_i^{(\pi,2)}(n) = ng^{(2)} + w_i^{(2)} \qquad \text{where} \tag{24}$$

$$g^{(2)} + w_i^{(2)} = s_i^{(2)}(f) + \sum_{j \in \mathcal{I}} p_{ij}(f) w_j^{(2)}, \quad s_i^{(2)}(f) = \sum_{j \in \mathcal{I}} p_{ij}(f)[(r_{ij} + w_j) - (g + w_i)]^2. \tag{25}$$

To establish the growth rate of $\widetilde{M}_i^{(\pi,3)}(n)$, it suffices to insert into (21) from (20). Since $\sum_{j\in\mathcal{I}} p_{ij}(f)\,[(r_{ij}+w_j)-(g+w_i)]\,g^{(2)}=0$ we can conclude

$$\sum_{j\in\mathcal{I}} p_{ij}(f)\,[(r_{ij}+w_j)-(g+w_i)]\,(ng^{(2)}+w_j^{(2)})=\sum_{j\in\mathcal{I}} p_{ij}(f)\,[(r_{ij}+w_j)-(g+w_i)]\,w_j^{(2)}.$$

Hence using the same arguments as for the second central moment we can conclude that

$$\widetilde{M}_i^{(\pi,3)}(n)=ng^{(3)}+w_i^{(3)}\quad\text{where}\quad g^{(3)}+w_i^{(3)}=s_i^{(3)}(f)+\sum_{j\in\mathcal{I}} p_{ij}w_j^{(3)} \tag{26}$$

$$s_i^{(3)}(f)=\sum_{j\in\mathcal{I}} p_{ij}(f)\left\{[(r_{ij}+w_j)-(g+w_i)]^3+3\,[(r_{ij}+w_j)-(g+w_i)]\,w_j^{(2)}\right\}. \tag{27}$$

4 Continuous-time Models: Exponential Utility and Higher Moments

Supposing that the obtained random reward up to time t, say $\xi(t)$, is evaluated by an exponential utility function, say $u^\gamma(\cdot)$, with the risk sensitivity coefficient γ, let for $\pi\sim(f)$, $U_i^{(\gamma)}(t,f):=\mathsf{E}_i^{\pi}[\exp(\gamma\xi(t))]$ considered as the moment generating function of $\xi(t)$, we can conclude that for $k=0,1,2,\ldots,n=0,1,2,\ldots$

$$M_i^{(k,\pi)}(t):=\mathsf{E}_i^{\pi}(\exp(\xi(t)^k)=\frac{\mathrm{d}^k}{\mathrm{d}\gamma^k}\mathsf{E}_i^{\pi}[\exp(\gamma\xi(t))]|_{\gamma=0}\quad\text{is the } k\text{th moment of } \xi(t) \tag{28}$$

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

$$U_i^{(\gamma)}(t,f)=1+\sum_{k=1}^{\infty}\frac{\gamma^k}{k!}M_i^{(k,\pi)}(t). \tag{29}$$

Similarly on introducing the moment generating function for the central moments of $\xi(t)$ by

$$\widetilde{U}_i^{(\gamma)}(t,f):=\mathsf{E}_i^{\pi}[\exp(\gamma(\xi(t)-\mathsf{E}_i^{\pi}\xi(t))]^k\ \text{ for all }\ i\in\mathcal{I} \tag{30}$$

for the kth central moments of $\xi(t)$ we have

$$\widetilde{M}_i^{(k,\pi)}(t):=\mathsf{E}_i^{\pi}[\xi(t)-\mathsf{E}_i^{\pi}\xi(t)]^k=\frac{\mathrm{d}^k}{\mathrm{d}\gamma^k}\mathsf{E}_i^{\pi}[\exp(\gamma(\xi(t)-\mathsf{E}_i^{\pi}\xi(t))]_{\gamma=0} \tag{31}$$

and the Taylor expansion around $\gamma=0$ for sufficiently small γ reads

$$\widetilde{U}_i^{(\gamma)}(t,f)=1+\sum_{k=1}^{\infty}\frac{\gamma^k}{k!}\widetilde{M}_i^{(k,\pi)}(t). \tag{32}$$

Let $M^{(k,\pi)}(t),\ \widetilde{M}^{(k,\pi)}(t)$ be the (column) vector of the k moments, central k moments respectively, with elements $M_i^{(k,\pi)}(t),\ \widetilde{M}_i^{(k,\pi)}(t)$ respectively.

In particular, if the system starts in state i, the expected total reward earned in state i up to the first exit of state i within time interval $[t,t+\delta)$ is equal to

$$M_i^{(1,\pi)}(t+\delta)=M_i^{(1,\pi)}(t)\cdot[1-q_i(f_i)\cdot\delta]+\delta\cdot r(i)\cdot[1-q_i(f_i)\cdot\delta]+o(\delta^2) \tag{33}$$

and for the s-th power of this reward it holds

$$\begin{aligned}M_i^{(s,\pi)}(t+\delta)&=\{M_i^{(1,\pi)}(t)[1-q_i(\cdot)\cdot\delta]+[r(i)\cdot\delta]\cdot[1-q_i(f_i)\cdot\delta]\}^s\\&=M_i^{(s,\pi)}(t)+s\cdot M_i^{(s-1,\pi)}(t)\cdot\delta+o(\delta^2)\end{aligned} \tag{34}$$

(Observe that $M_i^{(s,\pi)}(t)=[M_i^{(1,\pi)}(t)]^s$, $[M_i^{(1,\pi)}(t)+r(i)\cdot\delta]^s=M_i^{(s,\pi)}(t)+s\cdot r(i)\cdot M_i^{(s-1,\pi)}(t)+o(\delta^2)$.)

On inserting from (28),(30),(33) after some algebra and for δ tending to null we arrive at differential equations similar to that of discrete time models. In particular,

$$\text{For } k=1:\ \frac{\mathrm{d}}{\mathrm{d}t}M_i^{(1,\pi)}(t)=r(i)+\sum_{j\in\mathcal{I},j\neq i}q_{ij}(f_i)\,r_{ij}+\sum_{j\in\mathcal{I}}q_{ij}(f_i)\,M_j^{(1,\pi)}(t). \tag{35}$$

$$\begin{aligned}\text{For } k=2:\ \frac{\mathrm{d}}{\mathrm{d}t}M_i^{(2,\pi)}(t)&=2\cdot M_i^{(1,\pi)}(t)\cdot r(i)+\sum_{j\in\mathcal{I},j\neq i}q_{ij}(f_i)\left\{[r_{ij}]^2+2\cdot r_{ij}\cdot M_j^{(1,\pi)}(t)\right\}\\&\quad+\sum_{j\in\mathcal{I}}q_{ij}(f_i)\cdot M_j^{(2,\pi)}(t).\end{aligned}$$

In general:

$$\frac{\mathrm{d}}{\mathrm{d}t}M_i^{(s,\pi)}(t) = s \cdot M_i^{(s-1,\pi)}(t) \cdot r(i) + \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\left\{\sum_{k=1}^{s}\left(\begin{array}{c} s \\ k \end{array}\right) \cdot M_i^{(s,\pi)}(t) \cdot \left[r_{ij}\right]^k M_j^{(s-k,\pi)}(t)\right\}$$
$$+\sum_{j\in\mathcal{I}} q_{ij}(f_i) \cdot M_j^{(s,\pi)}(t). \tag{36}$$

Supposing that higher moments are known, the corresponding central moments can be easily computed. To this end, on recalling definition central moments, we can easily conclude that the if the system starts in state i and policy π is followed then the nth central moment at time t

$$\widetilde{M}_i^{(n,\pi)}(t) := \mathsf{E}_i^{\pi}[\xi(t) - M_i^{(1,\pi)}(t)]^n \tag{37}$$

Since $M_i^{(j,\pi)}(t) := \mathsf{E}_i^{\pi}[\xi(t)]^j$, after little algebra we arrive at

$$\widetilde{M}_i^{(n,\pi)}(t) := \sum_{j=0}^{n}\left(\begin{array}{c} n \\ j \end{array}\right) \cdot (-1)^{n-j} \cdot M_i^{(j,\pi)}(t) \cdot [M_i^{(1,\pi)}(t)]^{n-j} \tag{38}$$

$$= \sum_{j=0}^{n-2}\left(\begin{array}{c} n \\ j \end{array}\right) \cdot (-1)^{n-j} \cdot M_i^{(j,\pi)}(t) \cdot [M_i^{(1,\pi)}(t)]^{n-j} + (-1)^{n-1} \cdot (n-1) \cdot [M_i^{(1,\pi)}(t)]^n. \tag{39}$$

More details can be found in Sladký (2020).

5 Conclusions

The paper presents explicit formulas for higher moments and higher central moments of cumulative rewards in Markov reward chains that can be applied for more sophisticated optimality criteria in many dynamic problems.

Acknowledgements: This work was supported by the Czech Science Foundation under Grant 18-02739S.

References

[1] Cavazos-Cadena, R. and Fernandez-Gaucherand, F. (1999). Controlled Markov ..Chains with Risk-Sensitive Criteria: Average Cost, Optimality Equations and Optimal Solutions. *Mathematical Methods of Operations Research*, 43, pp. 121–139.

[2] Cavazos-Cadena, R. and Hernández-Hernández, D. (2005). A Characterization of the Optimal Risk-Sensitive Average Cost Infinite Controlled Markov Chains. *Annals of Applied Probability*, 15, pp. 175–212.

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

[4] Howard, R. A. and Matheson, J. (1972). Risk-Sensitive Markov Decision Processes. *Management Science*, 18(7), pp. 356–369.

[5] Jaquette, S. C. (1973). Markov Decision Processes with a New Optimality Criterion: Discrete Time. *Annals of Statistics*, 1, pp. 496–505.

[6] Mandl, P. (1971). On the Variance in Controlled Markov Chains. *Kybernetika*, 7(1), pp. 1–12.

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

[8] Prieto–Rumeau, T. and Hernández-Lerma, O. (2009). Variance Minimization and the Overtaking Optimality Approach in Continuous-Time Controlled Markov Chains. *Mathematical Methods of Operations Research*, 70, pp. 527–540.

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

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

[11] Sladký, K. (2005). On Mean Reward Variance in Semi-Markov Processes. *Mathematical Methods of Operations Research*, 62(3), pp. 387–397

[12] Sladký, K. (2008). Growth Rates and Average Optimality in Risk-Sensitive Markov Decision Chains. *Kybernetika*, 44(2), pp. 205–216.

[13] Sladký, K. (2018). *Central Moments and Risk-Sensitive Optimality in Markov Reward Chains.* In: Quantitative Methods in Economics – Multiple Criteria Decision Making XIX (M. Reiff, P. Gežík, Eds). University od Economics, Bratislava 2018, pp. 325–331.

[14] Sladký, K. (2020). *Central Moments and Risk-Sensitive Optimality in Continuous-Time Markov Reward Processes.* In: Quantitative Methods in Economics – Multiple Criteria Decision Making XX (M. Reiff, P. Gežík, Eds). University od Economics, Bratislava 2020, pp. 305–311.

[15] Sobel, M. (1982). The Variance of Discounted Markov Decision Processes. *Journal of Applied Probability*, 19, pp. 794–802.

[16] van Dijk, N. M. and Sladký, K. (2006). On the Total Reward Variance for Continuous-Time Markov Reward Chains. *Journal of Applied Probability*, 43(4), pp. 1044–1052.