Risk-Sensitivity and Average Optimality in Markov and Semi-Markov Reward Processes

Karel Sladký [1]

Abstract. This contribution is devoted to risk-sensitivity in long-run average optimality of Markov and semi-Markov reward processes. Since the traditional average optimality criteria cannot reflect the variability-risk features of the problem, we are interested in more sophisticated approaches where the stream of rewards generated by the Markov chain that is evaluated by an exponential utility function with a given risk sensitivity coefficient. Recall that for the risk sensitivity coefficient equal to zero (i.e. the so called risk-neutral case) we arrive at traditional optimality criteria, if the risk sensitivity coefficient is close to zero the Taylor expansion enables to evaluate variability of the generated total reward. Observe that the first moment of the total reward corresponds to expectation of total reward and the second central moment to the reward variance. In this note we present necessary and sufficient risk-sensitivity and risk-neutral optimality conditions for long run risk-sensitive average optimality criterion of unichain Markov and semi-Markov reward processes.

Keywords: Markov and semi-Markov reward processes, exponential utility function, risk sensitivity, long run optimality.

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

1 Formulation and Notation

Consider a controlled semi-Markov reward process $Y = \{Y(t), t \geq 0\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$ along with the embedded Markov chain $X = \{X_n, n = 0, 1, \ldots\}$. The development of the process $Y(t)$ over time is the following: At time $t = 0$ if $Y(0) = i$ the decision maker selects decision from an infinite (compact) set $\mathcal{A}_i \equiv [0, K_i] \subset \mathbb{R}$ of possible decisions (actions) in state $i \in \mathcal{I}$. Then state j is reached in the next transition with a given probability $p_{ij}(a)$ after random time $\eta_{ij}(a)$. Let $F_{ij}(a, \tau)$ be a non-lattice distribution function representing the conditional probability $\mathrm{P}(\eta_{ij} \leq \tau)$. We assume that for $\ell = 1, 2, \quad 0 < d_{ij}^{(\ell)} = \int_0^\infty \tau^\ell \, \mathrm{d}F_{ij}(a, \tau) < \infty$ hence also $0 < d_i^{(\ell)} = \sum_{j=1}^N p_{ij}(a) d_{ij}^{(\ell)}(a) < \infty$. Finally, one-stage transition reward $r_{ij} > 0$ will be accrued to transition from state i to state j, and reward rate $r_i(a)$ per unit of time incurred in state i is earned. We assume that each $p_{ij}(a)$ and $r_i(a)$ is a continuous function of $a \in \mathcal{A}_i$.

A (Markovian) policy controlling the semi-Markov process Y, say $\pi = (f^0, f^1, \ldots)$, 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 nth transition if the embedded Markov chain X is in state i. Let π^k be a sequence of decision vectors starting at the k-th transition, hence $\pi = (f^0, f^1, \ldots f^{k-1}, \pi^k)$. Policy which selects at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary; $P(f)$ is transition probability matrix with elements $p_{ij}(f_i)$. Stationary policy $\tilde{\pi}$ is randomized if there exist decision vectors $f^{(1)}, f^{(2)}, \ldots, f^{(m)} \in \mathcal{F}$ and on following policy $\tilde{\pi}$ we select in state i action $f_i^{(j)}$ with a given probability $\kappa_i^{(j)}$ (of course, $\kappa_i^{(j)} \geq 0$ with $\sum_{j=1}^N \kappa_i^{(j)} = 1$ for all $i \in \mathcal{I}$). For details see e.g. [1, 9, 10].

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

For the (random) reward earned up to time t, say $\xi(t)$ we have $\xi(t) := \left[\int_0^t r_{Y(s)} \mathrm{d}s \, + \sum_{k=0}^{N(t)} r_{Y(\tau_k^-), Y(\tau_k^+)} \right]$, with $Y(s)$, denoting the state of the system at time s, $Y(\tau_k^-)$ and $Y(\tau_k^+)$ the state just prior and after the kth jump, $N(t)$ the number of jumps up to time t, and $\mathcal{R}_i(f, t) := \mathsf{E}_i^\pi \xi(t)$ denote the expected total reward of the semi-Markov process $Y(t)$ up to time t given its initial state at time $t = 0$ if policy $\pi \sim (f)$ is followed.

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

In this note, we assume that the stream of rewards generated by the Markov processes is evaluated by an exponential utility function, say $\bar{u}^\gamma(\cdot)$, i.e. a separable utility function with constant risk sensitivity $\gamma \in \mathbb{R}$. For more details see e.g. [2, 3, 4, 7, 14]. Then the utility assigned to the (random) outcome ξ is given by

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

Observe that exponential utility function $\bar{u}^\gamma(\cdot)$ is separable, and multiplicative if the risk sensitivity $\gamma \neq 0$ or 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 $\gamma \neq 0$ and $u^\gamma(\xi_1 + \xi_2) \equiv \xi_1 + \xi_2$ for $\gamma = 0$.

The certainty equivalent corresponding to ξ, say $Z^\gamma(\xi)$, is given by

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

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

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

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

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

In what follows let

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

In this note we focus attention on risk-neutral and risk-sensitive optimality of so called unichain models, i.e. when the underlying Markov chain contains a single class of recurrent state and present characterization of control policies by discrepancy functions. Discrepancy functions were originally introduced in [8] for risk-neutral unichain models, possible extension to multichain case can be found in [11, 12]. To this end we make

Assumption GA. 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 GA 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).

2 Risk-Neutral Optimality in Unichain Semi-Markov Processes

At first we focus attention on the embedded Markov chains and slightly extend some results reported in [13]. To this end, on introducing for arbitrary $g, w_j \in \mathbb{R}$ $(i, j \in \mathcal{I})$ and decision $f \in \mathcal{F}$, the discrepancy functions

$$\tilde{\varphi}_{i,j}(w^{\mathrm{c}}, g^{\mathrm{c}}, f) := d_i(f_i) \cdot r(i) + r_{ij} - w_i^{\mathrm{c}} + w_j^{\mathrm{c}} - g^{\mathrm{c}}, \qquad \bar{\varphi}_{i,j}(w^{\mathrm{t}}, g^{\mathrm{t}}, f) := d_i(f_i) - w_i^{\mathrm{t}} + w_j^{\mathrm{t}} - g^{\mathrm{t}} \tag{6}$$

for the random reward obtained, resp. time elapsed, up to the nth transition we have

$$\xi_n = ng^{\mathrm{c}} + w_{X_0}^{\mathrm{c}} - w_{X_n}^{\mathrm{c}} + \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w^{\mathrm{c}}, g^{\mathrm{c}}, f) \ \text{ resp. } \ \eta_n = ng^{\mathrm{t}} + w_{X_0}^{\mathrm{t}} - w_{X_n}^{\mathrm{t}} + \sum_{k=0}^{n-1} \bar{\varphi}_{X_k, X_{k+1}}(w^{\mathrm{t}}, g^{\mathrm{t}}, f). \tag{7}$$

Hence by (7) for the expectation of ξ_n, $\mathsf{E}_i^{\,\pi}\xi_n =: v_i^\pi(n)$, resp. of η_n, with $\mathsf{E}_i^{\,\pi}\eta_n =: t_i^\pi(n)$, we get

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

$$t_i^\pi(n) = ng^{\mathrm{t}} + w_i^{\mathrm{t}} + \mathsf{E}_i^{\,\pi}\{\sum_{k=0}^{n-1} \bar{\varphi}_{X_k, X_{k+1}}(w^{\mathrm{t}}, g^{\mathrm{t}}, f) - w_{X_n}^{\mathrm{t}}\}. \tag{9}$$

Now we show how to express average reward generated by the semi-Markov process $Y(t), t \geq 0$ in terms of the embedded Markov chain X_n. Considering policy $\pi \sim (f)$, let

$$\tilde{\varphi}_i(w^{\mathrm{c}}, g^{\mathrm{c}}, f) := \sum_{j\in\mathcal{I}} p_{ij}(f_i)\tilde{\varphi}_{i,j}(w^{\mathrm{c}}, g^{\mathrm{c}}, f) = \sum_{j\in\mathcal{I}} p_{ij}(f_i)[d_i(f_i)\cdot r(i) + r_{ij} - w_i^{\mathrm{c}} + w_j^{\mathrm{c}} - g^{\mathrm{c}}], \quad (10)$$

$$\bar{\varphi}_i(w^{\mathrm{t}}, g^{\mathrm{t}}, f) := \sum_{j\in\mathcal{I}} p_{ij}(f_i)\bar{\varphi}_{i,j}(w^{\mathrm{t}}, g^{\mathrm{t}}, f) = \sum_{j\in\mathcal{I}} p_{ij}(f_i)[d_i(f_i) - w_i^{\mathrm{t}} + w_j^{\mathrm{t}} - g^{\mathrm{t}}] \quad (11)$$

It is well-known from the dynamic programming literature (cf. e.g. [1, 6, 9, 10]) that for every $f \in \mathcal{F}$ and arbitrary transition costs $s_{ij} = d_i(f_i)r(i) + r_{ij}, i, j, \in \mathcal{I}$, there exist numbers $g(f)$ and $w_i(f), i \in \mathcal{I}$ (unique up to additive constant) such that

$$w_i(f) + g(f) = d_i(f_i)r(i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} + w_j(f)], \ \ (i \in \mathcal{I}), \quad \text{i.e.} \quad (12)$$

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

In particular, for suitable selected $w_j^{\mathrm{c}}(f)$, resp. $w_j^{\mathrm{t}}(f)$, we have

$$v_i^{\pi}(n) = ng^{\mathrm{c}}(f) + w_i^{\mathrm{c}}(f) - \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot w_j^{\mathrm{c}}(f), \qquad t_i^{\pi}(n) = ng^{\mathrm{t}} + w_i^{\mathrm{t}} - \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot w_j^{\mathrm{t}}(f), \quad \text{where} \quad (13)$$

$$w_i^{\mathrm{c}}(f)+g^{\mathrm{c}}(f) = d_i(f_i)\cdot r(i)+\sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij}+w_j^{\mathrm{c}}(f)], \text{ resp. } w_i^{\mathrm{t}}(f)+g^{\mathrm{t}}(f) = d_i(f_i)+\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot w_j^{\mathrm{t}}(f), \quad (i \in \mathcal{I}). \quad (14)$$

After some manipulation we obtain from (13)

$$w_i^{\mathrm{t}}(f)\cdot\frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)} + g^{\mathrm{c}}(f) = d_i(f_i)\cdot\frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)} + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot w_j^{\mathrm{t}}(f)\frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)} \quad (15)$$

and by subtracting (15) from (14) we get

$$w_i(f) = \bar{r}_i(f) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)w_j(f) - d_i(f_i)g(f) \qquad \text{where} \quad (16)$$

$$w_i(f) := w_i^{\mathrm{c}}(f) - w_i^{\mathrm{t}}(f)\cdot\frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)}, \quad g(f) := \frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)}, \quad \bar{r}_i(f) = d_i(f_i)\cdot r(i) + \sum_{j\in\mathcal{I}, j\neq i} p_{ij}(f_i)r_{ij}.$$

On introducing matrix notations $P(f) = [p_{ij}(f_i)]$, $D(f) = \mathrm{diag}\,[d_i(f_i)]$, (square matrices)
$\bar{r}(f) = [\bar{r}_i(f)]$, $w(f) = [w_i(f)]$, $\bar{g}(f) = [g(f)]$ (column vectors) equation (16) can be written as

$$w(f) = \bar{r}(f) + P(f)w(f) - D(f)g(f) \Rightarrow g(f) = D^{-1}(f)\bar{r}(f) + [D^{-1}(f)P(f) - I]\cdot w(f)]. \quad (17)$$

Let

$$\tilde{r}(f) := D^{-1}(f)\bar{r}(f), \qquad \tilde{w}(f) := D^{-1}(f)w(f), \qquad \tilde{P}(f) := D^{-1}(f)\cdot P(f)\cdot D(f)$$

Then for the elements of $\tilde{r}(f),\ \tilde{w}(f),\ \tilde{P}(f)$ we have

$$\tilde{r}_i(f) = \bar{r}(i) + [d_i(f_i)]^{-1}r_{ij}, \quad \tilde{p}_{ij}(f_i) := p_{ij}(f_i)\frac{[d_j(f_j)]}{[d_i(f_i)]}, \quad \tilde{w}_i(f) := [d_i(f_i)]^{-1}w_i(f).$$

In particular, let us consider continuous-time Markov decision chain with transition intensities $\mu_{ij}(f_i)$, where $\sum_{j\in\mathcal{I}, j\neq i}\mu_{ij}(f_i) = -\mu_{ii}(f_i)$ and $\mu_i(f_i) = -\mu_{ii}(f_i)$ is the is the intensity of jumps from state i. Obviously, this is a very special case of semi-Markov processes with transition probabilities $p_{ij}(f) = \frac{\mu_{ij}(f_i)}{\mu_i(f_i)}$, and expected holding time $d_i(f_i) = \frac{1}{\mu_i(f_i)}$ in state i. Then on replacing in (17) transition probabilities and expected holding times by transition intensities for the average reward per unit of time of the considered continuous-time Markov process we conclude that

$$g(f) = r(i) + \sum_{j\neq i}\mu_{ij}(f_i)r_{ij} + \sum_j \mu_{ij}(f_i)w_{ij}(f) \quad (18)$$

the standard equation for average reward of a continuous time Markov reward chain (cf. e.g. [6]).

3 Risk-Sensitive Optimality in Unichain Semi-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 GA is fulfilled. We show how the discrepancy functions can be employed for finding optimality conditions for risk-sensitive Markov and semi-Markov processes. These results slightly extend some previous results reported in [12, 14, 15, 16, 17].

Similarly to the risk-neutral models, let for real g, w_i's ($i \in \mathcal{I}$)

$$\varphi_{ij}(w,g,f) := r_{ij} + d_i(f)\cdot[r(i)-g] + w_j - w_i, \;\; \text{where} \;\; w' = \min_{i\in\mathcal{I}} w_i, \;\; w'' = \max_{i\in\mathcal{I}} w_i. \tag{19}$$

Then if policy $\pi = (f^n)$ is followed we get by (5),(19) for the risk-sensitive case

$$U_i^{\pi}(\gamma,n) = \mathsf{E}_i^{\pi} \mathrm{e}^{\gamma\sum\limits_{k=0}^{n-1}[d_{X_k}(f_{X_k})\cdot r(X_k) + r_{X_k,X_{k+1}}]} = \mathrm{e}^{\gamma w_i} \times \mathsf{E}_i^{\pi} \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1}\varphi_{X_k,X_{k+1}}(w,g,f) - w_{X_n}]}. \tag{20}$$

Hence for a given $\gamma \neq 0$ there exist numbers $\bar{w}, \tilde{w}$ such that for any policy $\pi = (f^n)$

$$\mathsf{E}_i^{\pi} \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1}\varphi_{X_k,X_{k+1}}(w,g,f) - \bar{w}]} \leq \frac{U_i^{\pi}(\gamma,n)}{\mathrm{e}^{\gamma w_i}} \leq \mathsf{E}_i^{\pi} \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1}\varphi_{X_k,X_{k+1}}(w,g,f) - \bar{w}]}. \tag{21}$$

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

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)\,\mathrm{e}^{\gamma[r_{ij}+d_i(f_i)r(i)+w_j(f)]} = \mathrm{e}^{\gamma[d_i(f_i)g(f)+w_i(f)]} \;\; \text{or} \;\; \mathsf{E}_i^{\pi}\,\mathrm{e}^{\gamma\,\varphi_{X_k,X_{k+1}}(w,g,f)} = 1. \tag{22}$$

Moreover, on introducing new variables

$$v_i(f) := \mathrm{e}^{\gamma w_i(f)}, \quad \rho(f) := \mathrm{e}^{\gamma g(f)}, \quad q_{ij}(f) := p_{ij}(f)\mathrm{e}^{\gamma[r_{ij}+d_i(f_i)r(i)]} \tag{23}$$

from (22) we arrive at the following set of equations

$$\sum_{j\in\mathcal{I}} q_{ij}(f_i)\,v_j(f) = \rho(f)^{[d_i(f_i)]}\cdot v_i(f). \qquad (i\in\mathcal{I}) \tag{24}$$

Observe that if all $d_i(f_i)$'s are equal to some constant, say d, then (24) is a well-known formula for finding spectral radius (or so called Perron eigenvalue) of a nonnegative matrix, $v_i(f)$'s are elements of the corresponding Perron eigenvector (cf. [5]). In particular, if for the all $i \in \mathcal{I}$ and $f \in \mathcal{F}$ the values $d_i(f_i)$'s are equal to one the considered semi-Markov reward process is reduced to a Markov reward chain and Eq.(24) to formulas for calculating γ-risk average reward/cost optimality equation of the Markov reward chain. Unfortunately, comparing with the risk-neutral model, unichain property itself (cf. Assumption GA), cannot guarantee positivity of the Perron eigenvector; however, Perron eigenvector is strictly positive if the respective matrix is irreducible.

In what follows we focus attention on finding stationary policies $\pi^* \sim (f^*)$, resp. $\hat{\pi} \sim (\hat{f})$, such that for any $f \in \mathcal{F}$ it holds $\rho(f^*) \geq \rho(f)$, resp. $\rho(\hat{f}) \leq \rho(f)$. We show that the policies $\pi^* \sim (f^*)$, resp. $\hat{\pi} \sim (\hat{f})$, can be found by policy iterations. To specify the policy iteration algorithm, it will be convenient to use the following matrix notation.[2]

On introducing the $N \times N$ matrices $Q(f) = [q_{ij}(f_i)]$ and (column) N-vector $v(f) = [v_i(f)]$ along with diagonal $N \times N$ matrix $B(f) = \mathrm{diag}[\rho(f)^{[d_i(f_i)]}]$, from (24) we get

$$B(f)\cdot v(f) = Q(f)\cdot v(f) \quad \Longleftrightarrow \quad v(f) = [B(f)]^{-1}\cdot Q(f)\cdot v(f). \tag{25}$$

Obviously, from (25) for the i-element of $v(f)$ it holds $v_i(f) = \rho(f)^{-d_i(f_i)}\cdot\sum_{j=1}^{N} q_{ij}(f_i)\cdot v_j(f)$.

If stationary policy $\pi \sim (f)$ is followed, policy improvement routine can be used for finding an improved decision in any state (such approach slightly extends policy iteration method reported in [7] for finding maximal possible spectral radius of a family of controlled ergodic Markov reward chains).

[2] In vector inequalities $a \geq b$ 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. Using matrix notations the symbol I is reserved for identity matrix, e denotes unit (column) vector.

a) $\sum_{j=1}^{N} q_{ij}(h_i) \cdot v_j(f) \geq v_i(f)$ if maximal $\rho(f)$ is seeking, resp.,

b) $\sum_{j=1}^{N} q_{ij}(h_i) \cdot v_j(f) \leq v_i(f)$ if minimal $\rho(f)$ is seeking.

Repeating the above procedure we can generate a sequence of stationary policies with increasing, resp. decreasing, sequence of the values $\rho(f)$'s converging to maximal, resp. minimal, value of $\rho(f)$'s.

To this end, since $Q^{(B)}(f) := [B(f)]^{-1} \cdot Q(f)$ is an irreducible nonnegative matrix, let for some $h \in \mathcal{F}$ $z^{(B)}(h)$ be the left Perron eigenvector of $Q^{(B)}(h)$, i.e. $z^{(B)}(h) \cdot Q^{(B)}(h) = \rho^{(B)}(h) \cdot z^{(B)}(h)$. Since the matrix $Q^{(B)}(\cdot)$ is irreducible, the vectors $z^{(B)}(h)$, $v(f)$ are strictly positive, hence also their product $z^{(B)}(h) \times v(f)$ is positive.

In particular, let

$$\psi(h,f) := [Q^{(B)}(h) - Q^{(B)}(f)] \cdot v(f), \qquad \bar{\psi}(h,f) := B(f) \cdot [Q^{(B)}(h) - Q^{(B)}(f)] \cdot v(f)$$

be column N-vectors with elements $\psi_i(h,f)$, resp. $\bar{\psi}_i(h,f)$. Then by (25) we can conclude that

$$Q^{(B)}(h) \cdot v(h) - Q^{(B)}(f) \cdot v(f) = Q^{(B)}(h) \cdot v(h) - Q^{(B)}(h) \cdot v(f) + \bar{\psi}(h,f)$$

and on premultiplying by $z^{(B)}(h)$ we can conclude that

$$[\rho^{(B)}(h) - \rho^{(B)}(f)] \cdot z^{(B)}(h) \cdot v(f) = z^{(B)}(h) \cdot \bar{\psi}(h,f).$$

Then $\bar{\psi}(h,f) > 0$ implies that $\rho^{(B)}(h) > \rho^{(B)}(f)$, if $\bar{\psi}(h,f) < 0$ then $\rho^{(B)}(h) < \rho^{(B)}(f)$.
Observe that if all $d_i(f)$'s are equal to one (or at least to some constant number) the value $\rho(f)$ in (25) is the Perron eigenvalue (equal to the spectral radius) of the nonnegative matrix $Q(f) = [q_{ij}(f)]$. To this end we can expect that for not too much different elements of the diagonal matrix $B(f)$ if $\bar{\psi}(h,f) > 0$, resp. $\bar{\psi}(h,f) < 0$, then also $\psi(h,f) > 0$, resp. $\psi(h,f) < 0$.

Repeating the above improvements, we arrive at stationary policies $\pi^* \sim (f^*)$, resp. $\hat{\pi} \sim (\hat{f})$, such that

$$v_i(f^*) = \max_{f \in \mathcal{F}} \sum_{j \in \mathcal{I}} q_{ij}(f_i) \, [\rho(f)]^{[-d_i(f_i)]} v_j(f_i) = \sum_{j \in \mathcal{I}} q_{ij}(f_i^*) \, [\rho(f^*)]^{[-d_i(f_i^*)]} v_j(f_i^*) \tag{26}$$

$$v_i(\hat{f}) = \min_{f \in \mathcal{F}} \sum_{j \in \mathcal{I}} q_{ij}(f_i) \, [\rho(f)]^{[-d_i(f_i)]} v_j(f_i) = \sum_{j \in \mathcal{I}} q_{ij}(\hat{f}_i) \, [\rho(\hat{f})]^{[-d_i(\hat{f}_i)]} \, v_j(\hat{f}_i) \tag{27}$$

Moreover, from (26), (27), we get the following set of nonlinear equations

$$\begin{aligned} \mathrm{e}^{\gamma w_i^*} &= \max_{f \in \mathcal{F}} \sum_{j \in \mathcal{I}} p_{ij}(f_i) \, \mathrm{e}^{\gamma [r_{ij} + d_i(f_i)[r(i) - g(f)] + w_j(f)]} \\ &= \mathrm{e}^{\gamma d_i(f_i^*)[r(i) - g(f^*)]} \sum_{j \in \mathcal{I}} p_{ij}(f_i^*) \, \mathrm{e}^{\gamma [r_{ij} + w_j(f^*)]} \quad (i \in \mathcal{I})) \end{aligned} \tag{28}$$

$$\begin{aligned} \mathrm{e}^{\gamma \hat{w}_i} &= \min_{f \in \mathcal{F}} \sum_{j \in \mathcal{I}} p_{ij}(f_i) \, \mathrm{e}^{\gamma [r_{ij} + d_i(f_i)[r(i) - g(f)] + w_j(f)]} \\ &= \mathrm{e}^{\gamma [d_i(\hat{f}_i)[r(i) - g(\hat{f})]} \sum_{j \in \mathcal{I}} p_{ij}(\hat{f}_i) \, \mathrm{e}^{\gamma [r_{ij} + w_j(\hat{f})]} \quad (i \in \mathcal{I}) \end{aligned} \tag{29}$$

Eqs. (28),(29) can be called the γ-*risk average reward/cost optimality equation* for semi-Markov processes.

Similar results can be also formulated for the corresponding certainty equivalents, see (2), (3). To this end, let $Z_i^{\pi}(\gamma, n) := \ln U_i^{\pi}(\gamma, n)$ hence in virtue of (20), (21) we can conclude from Eqs. (28),(29) that for $\hat{\pi} \sim (\hat{f})$, $\pi^* \sim (f^*)$ and any $\pi = (f^n)$

$$Z_i^{\pi}(\gamma, n) = \mathsf{E}_i^{\pi} \sum_{k=0}^{n-1} [d_{X_k}(f_{X_k}) \cdot r(X_k) + r_{X_k, X_{k+1}}] = \sum_{j \in \mathcal{I}} q_{ij}(f_i) \, [\rho(f)]^{[-d_i(f_i)]} \, v_j(f_i) \tag{30}$$

$$Z_i^{\hat{\pi}}(\gamma, n) \leq Z_i^{\pi}(\gamma, n) \leq Z_i^{\pi^*}(\gamma, n) \tag{31}$$

Special case: continuous-time Markov chain.

Let us consider continuous-time Markov decision chain with transition intensities $\mu_{ij}(f_i)$, for $i, j \in \mathcal{I}$, $j \neq i$, $\sum_{j\in\mathcal{I},j\neq i} \mu_{ij}(f_i) = -\mu_{ii}(f_i)$ and $\mu_i(f_i) = -\mu_{ii}(f_i)$ is the intensity of jumps from state i. Obviously, this is a very special case of semi-Markov processes with transition probabilities $p_{ij}(f_i) = \frac{\mu_{ij}(f_i)}{\mu_i(f_i)}$, for $j\neg i$, $p_{ii}(f_i) = 0$ and expected holding time $d_i(f_i) = \frac{1}{\mu_i(f_i)}$ in state i. Then by (22) for the considered continuous-time Markov process after some manipulation we conclude that

$$\sum_{j\in\mathcal{I},j\neq i} \mu_{ij}(f_i)\, \mathrm{e}^{\gamma[r_{ij}+\frac{1}{\mu_i(f_i)}r(i)+w_j(f)]} = \mu_i(f_i)\cdot \mathrm{e}^{\gamma[\frac{1}{\mu_i(f_i)}g(f)+w_i(f)]}. \tag{32}$$

Moreover, on introducing new variables (recall that $d_i(f_i) = \frac{1}{\mu_i(f_i)}$)

$$\bar{v}_i(f) := \mathrm{e}^{\gamma w_i(f)}, \quad \bar{\rho}(f) := \mathrm{e}^{\gamma g(f)}, \quad \bar{q}_{ij}(f) := \mu_{ij}(f)\mathrm{e}^{\gamma[r_{ij}+d_i(f_i)r(i)]} \tag{33}$$

from (32) we arrive at the following set of equations

$$\sum_{j\in\mathcal{I},j\neq i} \bar{q}_{ij}(f_i)\,\bar{v}_j(f) = \bar{\rho}(f)^{[d_i(f_i)]}\cdot \bar{v}_i(f) \qquad (i\in\mathcal{I}). \tag{34}$$

Acknowledgements

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

References

[1] Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control, Volume 2, Third Edition.* Belmont, Mass.: Athena Scientific.

[2] Cavazos-Cadena, R. & Montes-de-Oca, R. (2003). The value iteration algorithm in risk-sensitive average Markov decision chains with finite state space, *Math. Oper. Res.*, 28, 752–756.

[3] Cavazos-Cadena, R. (2003). 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, 253–285.

[4] Cavazos-Cadena, R. & Hernández-Hernández, D. (2005). A characterization of the optimal risk-sensitive average cost infinite controlled Markov chains, *Ann. Appl. Probab.,* 15, 175–212.

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

[6] Howard, R. A. (1960). *Dynamic Programming and Markov Processes.* Cambridge, Mass.: MIT Press.

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

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

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

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

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

[12] Sladký, K. (1980). Bounds on discrete dynamic programming recursions I, *Kybernetika*, 16, 526–547.

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

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

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

[16] Sladký, K. (2018). Risk-sensitive average optimality in Markov decision processes, *Kybernetika*, 54, 1218–1230.

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