Czech Economic Review 7 (2013) 146–161
Acta Universitatis Carolinae Oeconomica

Risk-Sensitive and Mean Variance Optimality in Markov Decision Processes

Karel Sladký*

Received 14 December 2012; Accepted 20 September 2013

Abstract In this paper we consider unichain Markov decision processes with finite state space and compact 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 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, that take into account not only the expected values of the total reward but also its higher moments.

Keywords Discrete-time Markov decision chains, exponential utility functions, certainty equivalent, mean-variance optimality, connections between risk-sensitive and risk-neutral models
JEL classification C44, C61

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. The best known approaches stem from the classical work of Markowitz (1952,1959) on mean variance selection rules, i.e. we optimize the weighted sum of the expected total (or average) reward and its variance.

On the other hand, risky decisions can be also eliminated when the generated random reward is evaluated by an exponential utility function and we optimize the corresponding expectation. To be more precise, let us consider an exponential utility function, say $\overline{u}^{\gamma}(\cdot)$, i.e. a separable utility function with constant risk sensitivity $\gamma \in \mathbb{R}$, where the utility assigned to the (random) outcome ξ is given by

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

* Academy of Sciences of the Czech Republic, Institute of Information Theory and Automation, Department of Econometrics, Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic. Phone: +420 266 052 366, E-mail: sladky@utia.cas.cz

For what follows let $u^\gamma(\xi) := \exp(\gamma\xi)$, hence $\overline{u}^\gamma(\xi) = (\text{sign}\gamma)u^\gamma(\xi)$. Recall that if $\gamma > 0$ (the risk seeking case) the decision maker prefers large values of ξ, if $\gamma < 0$ (the risk averse case) the decision maker prefers small values of ξ to the large ones.

Considering the utility function $\overline{u}^\gamma(\xi)$ the corresponding certainty equivalent, say $Z^\gamma(\xi)$, is the value such that $\overline{u}^\gamma(Z^\gamma(\xi)) = \mathrm{E}[\overline{u}^\gamma(\xi)]$ (the symbol E is reserved for expectation). Then we immediately get

$$Z^\gamma(\xi) = \begin{cases} \gamma^{-1}\ln\{\mathrm{E}u^\gamma(\xi)\}, & \text{if } \gamma \neq 0 \\ \mathrm{E}[\xi] & \text{for } \gamma = 0. \end{cases} \tag{2}$$

Recall that exponential utility functions are separable and hence suitable for sequential decisions in stochastic dynamic systems evolving over time.

The study of controlled Markov reward processes with exponential utility functions, called risk-sensitive optimality, was initiated in the seminal paper Howard and Matheson (1972). Recently, there has been an intensive work on Markov reward chain with risk sensitive optimality criteria (see e.g. Jaquette 1976; Hernández-Hernández and Marcus 1999; Borkar and Meyn 2002; Cavazos-Cadena 2002, 2003; Cavazos-Cadena and Fernández-Gaucherand 1999, 2000; Cavazos-Cadena and Montes de Oca 2003, 2005; Cavazos-Cadena and Hernández-Hernández 2002, 2004, 2005; Di Massi and Stettner 1999, 2000, 2006, 2007; and Sladký 2008a, 2008b, 2012).

In the present paper we restrict attention on unichain models with finite state space. At first we rederive optimality conditions for standard risk-neutral models with mean variance optimality and specify difficulties arising with this optimality criterion. Then we focus attention on the risk sensitivity and establish necessary and sufficient optimality conditions if the underlying Markov chain is irreducible or at least unichain and the risk sensitivity coefficient is sufficiently close to zero.

The paper is organized as follows. Section 2 contains notation and preliminaries, optimality equations for mean reward in risk-neutral (unichain) Markov processes are given in Section 3 and in Section 4 for the corresponding mean reward variance. Optimality equations for the risk-sensitive unichain models can be found in Section 5 and necessary and sufficient optimality conditions both for risk-sensitive and risk-neutral cases are presented in Section 6. Conclusions are made in Section 7. Some more technical proofs are presented in the Appendix.

2. Notation and Preliminaries

In this note, we consider Markov decision processes with finite state and compact action spaces where the stream of rewards generated by the Markov processes is evaluated by an exponential utility function (so-called risk-sensitive model) with a given risk sensitivity coefficient.

To this end, we consider at discrete time points Markov decision process $X = \{X_n,\ n = 0,1,\ldots\}$ with finite state space $\mathcal{I} = \{1,2,\ldots,N\}$, and compact set $\mathcal{A}_i = [0,K_i]$ 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 chosen, then state j is reached in the next transition with a given probability $p_{ij}(a)$

(depending continuously on a) and one-stage transition reward r_{ij} will be accrued to such transition.

A (Markovian) policy controlling the decision process is given by a sequence of decisions at every time point. In particular, policy controlling the process, $\pi = (f^0, f^1, \ldots)$, is identified by a sequence of decision vectors $\{f^n,\ n = 0,1,\ldots\}$ where $f^n \in \mathcal{F} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$ for every $n = 0,1,2,\ldots$, and $f_i^n \in \mathcal{A}_i$ is the decision (or action) taken at the n-th transition if the chain X is in state i. Let $\pi^m = (f^m, f^{m+1}, \ldots)$, hence $\pi = (f^0, f^1, \ldots, f^{m-1}, \pi^m)$, in particular $\pi = (f^0, \pi^1)$. The symbol E_i^π denotes the expectation if $X_0 = i$ and policy $\pi = (f^n)$ is followed, in particular, $\mathrm{E}_i^\pi(X_m = j) = \sum_{i_j \in \mathcal{I}} p_{i,i_1}(f_i^0) \ldots p_{i_{m-1},j}(f_{m-1}^{m-1})$; $\mathrm{P}(X_m = j)$ is the probability that X is in state j at time m.

Policy π which selects at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary, hence X is a homogeneous Markov chain with transition probability matrix $P(f)$ whose ij-th element equals $p_{ij}(f_i)$; $\mathrm{E}_i^\pi(X_m) = [P^m(f)]_{ij}$ (symbol $[A]_{ij}$ denotes the ij-th element of the matrix A) and $r_i(f_i) := \sum_{j \in \mathcal{I}} p_{ij}(f_i) r_{ij}$ is the expected reward obtained in state i. Similarly, $r(f)$ is an N-column vector of one-stage rewards whose i-the elements equals $r_i(f_i)$. The symbol I denotes an identity matrix and e is reserved for a unit column vector.

Recall that $P^*(f) := \lim_{n\to\infty} \frac{1}{n} \sum_{k=}^{n-1} P^k(f)$ (with elements $p_{ij}^*(f)$) exists, and if $P(f)$ is aperiodic then even $P^*(f) = \lim_{k\to\infty} P^k(f)$ and the convergence is geometrical. Moreover, if $P(f)$ is unichain, i.e. $P(f)$ contains a single class of recurrent states, then $p_{ij}^*(f) = p_j^*(f)$, i.e. limiting distribution is independent of the starting state.

We shall assume that the stream of transition rewards generated by the considered Markov decision process is evaluated by an exponential utility function (1) either with the risk aversion coefficient $\gamma \neq 0$ (the risk sensitive case) or with $\gamma = 0$ (the risk neutral case). To this end, let

$$\xi_n(\pi) = \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}$$

be the (random) total reward received in the n next transitions of the considered Markov chain X if policy $\pi = (f^n)$ is followed.

Supposing that $X_0 = i$, on taking expectation we have

$$\overline{U}_i^\gamma(\pi, n) := \mathrm{E}_i^\pi(\overline{u}^\gamma(\xi_n)) = (\mathrm{sign}\gamma)\mathrm{E}_i^\pi \mathrm{e}^{\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}} \qquad \text{if } \gamma \neq 0, \tag{3}$$

$$V_i(\pi, n) := \mathrm{E}_i^\pi(\overline{u}^\gamma(\xi_n)) = \mathrm{E}_i^\pi(\xi_n(\pi)) = \mathrm{E}_i^\pi \sum_{k=0}^{n-1} r_{X_k, X_{k+1}} \qquad \text{if } \gamma = 0, \tag{4}$$

$$S_i(\pi, n) := \mathrm{E}_i^\pi(\overline{u}^\gamma(\xi_n))^2 = \mathrm{E}_i^\pi(\xi_n(\pi))^2 = \mathrm{E}_i^\pi(\sum_{k=0}^{n-1} r_{X_k, X_{k+1}})^2 \qquad \text{if } \gamma = 0, \tag{5}$$

hence if $\gamma = 0$

$$\sigma_i(\pi, n) := \mathrm{E}_i^\pi[\xi_n - V_i(\pi, n)]^2 = S_i(\pi, n) - [V_i(\pi, n)]^2. \tag{6}$$

Observe that $\overline{U}_i^{\gamma}(\pi,n)$ is the expected utility, $V_i(\pi,n)$ and $S_i(\pi,n)$ is the first and second moment of the random variable $\xi_n(\pi)$ if the process starts in state i and $\sigma_i(\pi,n)$ is the corresponding variance.

If policy $\pi \sim (f)$ is stationary, the process X is time homogeneous and for $m < n$ we write $\xi_n = \xi_m + \xi_{n-m}$ (along with $X_0 = i$ we tacitly assume that $\mathrm{P}(X_m = j)$, hence ξ_{n-m} starts in state j). Then $[\xi_n]^2 = [\xi_m]^2 + [\xi_{n-m}]^2 + 2 \cdot \xi_m \cdot \xi_{n-m}$ and on taking expectations for $n > m$ we can conclude that

$$\begin{aligned} \mathrm{E}_i^{\pi}[\xi_n] &= \mathrm{E}_i^{\pi}[\xi_m] + \mathrm{E}_i^{\pi}\{\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j) \cdot \mathrm{E}_j^{\pi}[\xi_{n-m}]\}. && (7)\\ \mathrm{E}_i^{\pi}[\xi_n]^2 &= \mathrm{E}_i^{\pi}[\xi_m]^2 + \mathrm{E}_i^{\pi}\{\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j) \cdot \mathrm{E}_j^{\pi}[\xi_{n-m}]^2\} \\ &\quad +2 \cdot \mathrm{E}_i^{\pi}[\xi_m] \sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j) \cdot \mathrm{E}_j^{\pi}[\xi_{n-m}]. && (8) \end{aligned}$$

From (7) and (8) we directly conclude for $m = 1$

$$\begin{aligned} \mathrm{E}_i^{\pi}[\xi_n] &= r_i^{(1)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot \mathrm{E}_j^{\pi}[\xi_{n-1}], && (9)\\ \mathrm{E}_i^{\pi}[\xi_n]^2 &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot \mathrm{E}_j^{\pi}[\xi_{n-1}]^2 + 2 \cdot \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot r_{ij} \cdot \mathrm{E}_j^{\pi}[\xi_{n-1}], && (10) \end{aligned}$$

where $r_i^{(1)}(f_i) = r_i(f_i) = \sum_{j\in\mathcal{I}} p_{ij}(f_i)\ r_{ij}$, $r_i^{(2)}(f_i) = \sum_{j\in\mathcal{I}} p_{ij}(f_i)[\ r_{ij}]^2$.
By using the more appealing notation $V_i(f,n) = \mathrm{E}_i^{\pi}[\xi_n]$, (9), (10) take on the forms:

$$\begin{aligned} V_i(f,n+1) &= r_i^{(1)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot V_j(f,n), && (11)\\ S_i(f,n+1) &= r_i^{(2)}(f_i) + 2\sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot\ r_{ij} \cdot V_j(f,n) + \sum_{j\in\mathcal{I}} p_{ij}(f_i) S_j(f,n), && (12) \end{aligned}$$

or in matrix form as:

$$\begin{aligned} V(f,n+1) &= r^{(1)}(f) + P(f) \cdot V(f,n), && (13)\\ S(f,n+1) &= r^{(2)}(f) + 2 \cdot P(f) \cdot R \cdot V(f,n) + P(f) S(f,n), && (14) \end{aligned}$$

where $R = [r_{ij}]_{i,j}$ is an $N \times N$-matrix and $r^{(2)}(f) = [\ r_i^{(2)}(f_i)]_i$, $S(f,n) = [S_i(f,n)]_i$ are column vectors.

Finally, on inserting from (12) in (6) we get for the variance $\sigma_i(f,n)$

$$\begin{aligned} \sigma_i(f,n+1) &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\sigma_j(f,n) + 2 \cdot \sum_{j\in\mathcal{I}} p_{ij}(f_i)\ r_{ij} \cdot V_j(f,n) \\ &\quad -[V_i(f,n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)[V_j(f,n)]^2 \\ &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\sigma_j(f,n) + 2 \cdot \sum_{j\in\mathcal{I}} p_{ij}(f_i)\ r_{ij} \cdot V_j(f,n) \\ &\quad - \sum_{j\in\mathcal{I}} p_{ij}(f_i)[V_i(f,n+1) + V_j(f,n)][V_i(f,n+1) - V_j(f,n)]. && (15) \end{aligned}$$

3. Risk-neutral case: optimality equations

To begin with, (cf. Mandl 1971, Sladký 1974) first observe that if the discrepancy function

$$\overline{\varphi}_{ij}(w,\overline{g}) := r_{ij} - \overline{g} + w_j - w_i, \text{ for arbitrary } \overline{g},\ w_i \in \mathbb{R},\ i,j \in \mathcal{I}, \tag{16}$$

then by (4), (11)

$$\begin{aligned} V_i(\pi,n) &= \overline{g} + w_i + \sum_{j\in\mathcal{I}} p_{ij}(f_i^0)\{\overline{\varphi}_{ij}(w,\overline{g}) + V_j(\pi^1,n-1) - w_j\} && (17)\\ &= n\overline{g} + w_i + \mathrm{E}_i^{\pi} \sum_{k=0}^{n-1} \overline{\varphi}_{X_k,X_{k+1}}(w,\overline{g}) - \mathrm{E}_i^{\pi} w_{X_n}. && (18) \end{aligned}$$

In what follows let $\varphi_i(f_i,w,\overline{g}) := \sum_{j=1}^{N} p_{ij}(f_i)\,\overline{\varphi}_{ij}(w,\overline{g}))$ be the expected discrepancy accrued to state $i \in \mathcal{I}$, and denote by $\varphi(f,w,\overline{g}))$ the corresponding N-dimensional column vector of expected discrepancies. Then $[P(f)]^n \cdot \varphi(f,w,\overline{g})$ is the (column) vector of expected discrepancies accrued after n transitions; its i-th entry denotes the discrepancy if the process X starts in state i, g is a constant vector with elements $\overline{g}$, w is a column vector with elements w_i.

Similarly, for the vector of total expected rewards earned up to the n-th transition we get

$$V(\pi,n) := \sum_{k=0}^{n-1}\prod_{j=0}^{k-1} P(f^j)r(f^k) = ng + w + \sum_{k=0}^{n-1}\prod_{j=0}^{k-1} P(f^j)\varphi(f^k,w,\overline{g}) - \prod_{j=0}^{k-1} P(f^j)w \tag{19}$$

and its i-th element $V_i(\pi,n)$ is the total expected reward if the process starts in state i. Observe that for $n \to \infty$ elements of $V(\pi,n)$ can be typically infinite.

Moreover, following stationary policy $\pi \sim (f)$ for n tending to infinity there exist vector of average expected rewards, denoted $g(f)$ (with elements $g_i(f)$), where

$$g(f) = \lim_{n\to\infty} \frac{1}{n} V(f,n) = P^*(\pi)\, r(f). \tag{20}$$

If $P(f)$ is unichain, then all rows of $P^*(\pi)$ are equal to $p^*(\pi)$, hence $g(f)$ is a constant vector with elements

$$\overline{g}(f) = p^*(\pi) r(f). \tag{21}$$

Assumption A. There exists state $i_0 \in \mathcal{I}$ that is accessible from any state $i \in \mathcal{I}$ for every $f \in \mathcal{F}$, i.e. for every $f \in \mathcal{F}$ the transition probability matrix $P(f)$ is *unichain*.

The following facts are well-known to specialists in stochastic dynamic programming (see e.g. Howard 1960; Puterman 1994; Ross 1983).

Theorem 1. *If Assumption A holds, then*

(i) *For every $f \in \mathcal{F}$ there exist numbers $\overline{g}(f)$, $w_i(f)$'s ($i \in \mathcal{I}$) such that*

$$\overline{\varphi}_i(f, w(f), \overline{g}(f)) = 0 \Leftrightarrow r_i(f_i) - \overline{g}(f) + \textstyle\sum_{i \in \mathcal{I}} p_{ij}(f_i) w_j(f) - w_i(f) = 0 \; \forall i \in \mathcal{I},$$

or in matrix form

$$\overline{\varphi}(f, w(f), \overline{g}(f)) = 0 \Leftrightarrow w(f) + g(f) = r(f) + P(f) w(f),$$

where $g(f) = P^*(f) r(f)$.

(ii) *There exists a decision vector $f^* \in \mathcal{F}$ (resp. $\hat{f} \in \mathcal{F}$) along with (column) vectors $w^* = w(f^*)$, $\hat{w} = w(\hat{f})$ with elements w_i^*, $\hat{w}_i$ respectively, and $g^* = g(f^*)$ (resp. $\hat{g} = g(\hat{f})$) (constant vector with elements $\overline{g}(f) = p^*(f) r(f)$) being the solution of the (nonlinear) equation (I denotes the identity matrix)*

$$\max_{f \in \mathcal{F}} \left[r(f) - g^* + (P(f) - I) \cdot w^* \right] = 0, \quad \min_{f \in \mathcal{F}} \left[r(f) - \hat{g} + (P(f) - I) \cdot \hat{w} \right] = 0, \tag{22}$$

where $w(f)$ for $f = f^*$, $\hat{f}$ is unique up to an additive constant, and unique under the additional normalizing condition $P^*(f)\, w(f) = 0$. Then for

$$\begin{aligned} \varphi(f, f^*) &:= r(f) - g(f^*) + (P(f) - I) \cdot w(f^*), \\ \varphi(f, \hat{f}) &:= r(f) - g(\hat{f}) + (P(f) - I) \cdot w(\hat{f}) \end{aligned} \tag{23}$$

we have $\varphi(f, f^*) \leq 0$, $\varphi(f, \hat{f}) \geq 0$ with $\varphi(f^*, f^*) = \varphi(\hat{f}, \hat{f}) = 0$.

In particular, by (22)–(23) for every $i \in \mathcal{I}$ we can write

$$\begin{aligned} \varphi_i(f, f^*) &= r_i(f_i) - \overline{g}^* + \textstyle\sum_{j \in \mathcal{I}} p_{ij}(f_i) w_j^* - w_i^* \leq 0, \\ \varphi_i(f, \hat{f}) &= r_i(f_i) - \hat{\overline{g}} + \textstyle\sum_{j \in \mathcal{I}} p_{ij}(f_i) \hat{w}_j - \hat{w}_i \geq 0. \end{aligned}$$

Finally, if stationary policy $\pi \sim (f)$ is followed, there exist $g(f)$, $w_i(f)$'s (for $i \in \mathcal{I}$) such that

$$\varphi_i(f, f) = r_i(f_i) - \overline{g}(f) + \sum_{j \in \mathcal{I}} p_{ij}(f_i) w_j(f) - w_i(f) = 0 \text{ with } \sum_{j \in \mathcal{I}} p_j^*(f_i) w_j(f) = 0$$

and hence by (19)

$$V(f, n) := P(f)^n r(f) = n g(f) + w(f) - P(f)^n w(f). \tag{24}$$

If $P(f)$ is aperiodic for n tending to infinity we get

$$\lim_{n \to \infty} P(f)^n = P^*(f), \text{ hence } P^*(f)\, w(f) = 0 \tag{25}$$

and the convergence is geometrical.

4. Risk-neutral case: mean reward variance

In this section we focus attention on the risk neutral case where the variability can be evaluated by the mean variance of total expected reward. To this end, we shall consider fixed stationary policy $\pi \sim (f)$. As we shall see under Assumption A for a given stationary policy $\pi \sim (f)$ also $\sigma_i(f,n)$ (variance of total expected reward) grows for $n \to \infty$ linearly over time and its growth rate (independent of the starting state) can be calculated similarly as the mean reward. Recall that for the considered stationary policy $\pi \sim (f)$ the mean reward $g(f) = p^*(f)r(f)$. Similar formula can be also used for calculating mean variance as it is stated in the following theorem (for the proof see the Appendix).

Theorem 2. *Let Assumption A hold. Then the growth rate of the variance $\sigma_i(f,n)$ over time is linear and independent of the starting state i, i.e.*

$$G(f) = \lim_{n\to\infty} \frac{\sigma_i(f,n)}{n} \quad for\ all\ \ i \in \mathcal{I}. \tag{26}$$

Moreover, there exists column vector $s(f)$ with elements $s_i(f)$ such that

$$G(f) = p^*(f)s(f), \tag{27}$$

where

$$\begin{aligned} s_i(f) &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\big\{[w_j(f)]^2 + 2\, r_{ij}(f_i)w_j(f)\big\} - [\overline{g}(f) + w_i(f)]^2 \\ &= \sum_{j\in\mathcal{I}} p_{ij}(f_i)[\, r_{ij}(f_i) + w_j(f)]^2 - [\overline{g}(f) + w_i(f)]^2 \qquad (28) \\ &= \sum_{j\in\mathcal{I}} p_{ij}(f_i)[\, r_{ij}(f_i) + w_j(f) - \overline{g}(f)]^2 - [w_i(f)]^2. \qquad (29) \end{aligned}$$

Comparing (21) and (27) the difference is only in the column vectors $r(f)$ and $s(f)$. Elements $r_i(f_i)$ depends only on the decision taken in state i, i.e. on transition probabilities $p_{ij}(f_i)$ and transition rewards r_{ij}, however elements $s_i(f)$ depends also on $\overline{g}(f)$ and $w_j(f)$ for all $j \in \mathcal{I}$. Considering stationary policies $\pi \sim (f)$ and $\pi' \sim (f')$ we can easily calculate the corresponding average rewards $g(f)$, $g(f')$, but for calculating the mean variances $G(f)$, $G(f')$ in virtue of (28), (29) it is necessary to find along with $g(f)$, $g(f')$ also the values $w_j(f)$, $w_j(f')$ for all $j \in \mathcal{I}$. The calculation can be simplified only if $w_j(f) = w_j(f')$ for all $j \in \mathcal{I}$, e.g. if there are two stationary policies maximizing or minimizing mean reward, then we can easily select optimal policy minimizing the mean variance. In some papers the problem is simplified by replacing in (27) $s_i(f)$ only by $r_i^{(2)}(f_i)$, but the resulting policy need not minimize the variance (see e.g. Filar et al. 1989; Huang and Kallenberg 1994; Kadota 1997; Kawai 1987; Kurano 1987; Mandl 1971; Sladký and Sitař 2004; Sladký 2005; and Sobel 1985).

On the other hand, for the risk sensitive optimality, expectation of the exponential utility function takes into account not only expected value of the (random) reward but also all its higher moments.

5. Risk-sensitive models: optimality equations

For the risk-sensitive models, let $U_i^\gamma(\pi,n) := \mathrm{E}_i^\pi(u^\gamma(\xi_n))$ and hence $Z_i^\gamma(\pi,n) = \frac{1}{\gamma}\ln U_i^\gamma(\pi,n)$ is the corresponding certainty equivalent. In analogy to (17), (18) for expectation of the utility function we get by (16) for arbitrary $\overline{g}$, $w_i \in \mathbb{R}$, $i,j \in \mathcal{I}$

$$\begin{aligned} U_i^\gamma(\pi,n) &= \mathrm{e}^{\gamma(\overline{g}+w_i)} \sum_{j\in\mathcal{I}} p_{ij}(f_i^0) e^{\gamma\{\overline{\varphi}_{ij}(w,\overline{g})-w_j\}} \cdot U_j^\gamma(\pi^1,n-1) \qquad (30)\\ &= \mathrm{e}^{\gamma(2\overline{g}+w_i)} \sum_{j\in\mathcal{I}}\sum_{k\in\mathcal{I}} p_{ij}(f_i^0) e^{\gamma\{\overline{\varphi}_{ij}(w,\overline{g})-w_j\}} \cdot \mathrm{e}^{\gamma w_j} p_{jk}(f_j^1) e^{\gamma\{\overline{\varphi}_{jk}(w,\overline{g})-w_k\}} \cdot \\ &\quad \cdot U_k^\gamma(\pi^2,n-2) \\ &\ \ \vdots \\ &= \mathrm{e}^{\gamma(n\overline{g}+w_i)} \mathrm{E}_i^\pi e^{\gamma\{\sum_{k=0}^{n-1}\overline{\varphi}_{X_k,X_{k+1}}(w,\overline{g})-w_{X_n}\}}. \qquad (31) \end{aligned}$$

In particular, for stationary policy $\pi \sim (f)$ assigning numbers $g(f)$, $w_i(f)$ by (16) we have

$$\overline{\varphi}_{ij}(w(f),\overline{g}(f)) := r_{ij} - \overline{g}(f) + w_j(f) - w_i(f) \qquad (32)$$

and (30), (31) take on the form

$$\begin{aligned} U_i^\gamma(f,n) &= \mathrm{e}^{\gamma(\overline{g}(f)+w_i(f))} \sum_{j\in\mathcal{I}} p_{ij}(f_i) e^{\gamma\{\overline{\varphi}_{ij}(w(f),\overline{g}(f))-w_j(f)\}} \cdot U_j^\gamma(f,n-1) \\ &= \mathrm{e}^{\gamma(n\overline{g}(f)+w_i(f))} \mathrm{E}_i^\pi e^{\gamma\{\sum_{k=0}^{n-1}\overline{\varphi}_{X_k,X_{k+1}}(w(f),\overline{g}(f))-w_{X_n}(f)\}}. \end{aligned}$$

In what follows we show that under certain assumptions there exist $w_i(f)$'s, $g(f)$ such that

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)\mathrm{e}^{\gamma r_{ij}} \cdot \mathrm{e}^{\gamma w_j(f)} = \mathrm{e}^{\gamma\overline{g}(f)} \cdot \mathrm{e}^{\gamma w_i(f)}, \text{ for } i \in \mathcal{I}. \qquad (33)$$

Now let $\rho(f) := \mathrm{e}^{\gamma g(f)}$, $z_i(f) := \mathrm{e}^{\gamma w_i(f)}$, $q_{ij}(f_i) := p_{ij}(f_i)\mathrm{e}^{\gamma r_{ij}}$ and introduce the following matrix notation for column vectors $U^\gamma(\pi,n) = [U_i^\gamma(\pi,n)]_i$, $z(f) = [z_i(f)]_i$ and an $N \times N$ nonnegative matrix $Q(f) = [q_{ij}(f_i)]_{ij}$.

Then by (33) for stationary policy $\pi \sim (f)$ we immediately have $\rho(f)z(f) = Q(f)z(f)$. Since $Q(f)$ is a nonnegative matrix by the well-known Perron-Frobenius theorem $\rho(f)$ equals the spectral radius of $Q(f)$ and $z(f)$ can be selected nonnegative. Moreover, if $P(f)$ is irreducible then $Q(f)$ is irreducible, and $z(f)$ can be selected strictly positive. Recall (cf. Gantmakher 1959) that $z(f)$ can be selected strictly positive if and only if for suitable labelling of states of the underlying Markov chain (i.e. on suitably permuting rows and corresponding columns of $Q(f)$) it is possible to decompose $Q(f)$ such that:

$$Q(f) = \begin{bmatrix} Q_{(\mathrm{NN})}(f) & Q_{(\mathrm{NB})}(f) \\ 0 & Q_{(\mathrm{BB})}(f) \end{bmatrix},$$

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

- $\rho_{(\mathrm{N})}(f) < \rho(f)$;

- $\rho_{(\mathrm{B})}(f) = \rho(f)$ and $Q_{(\mathrm{BB})}(f)$ is block-diagonal, in particular,

$$Q_{(\mathrm{BB})}(f) = \begin{bmatrix} Q_{(11)}(f) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Q_{(rr)}(f) \end{bmatrix},$$

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

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

Finally observe that if $P(f)$ is unichain then $z(f)$ can be selected strictly positive if the risk sensitivity coefficient γ is sufficiently close to zero.

In (33) attention is focused only on a fixed stationary policy $\pi \sim (f)$. The above facts can be extended to all feasible policies under the following:

Assumption B. There exists state $i_0 \in \mathcal{I}$ that for every $f \in \mathcal{F}$ i_0 is accessible from any state $i \in \mathcal{I}$, i.e. for every $f \in \mathcal{F}$ the transition probability matrix $P(f)$ is unichain. Furthermore, if for some $f \in \mathcal{F}$ the matrices $P(f)$ and also $Q(f)$ are reducible then state i_0 belongs to the basic class of $Q(f)$ (observe that each $Q(f)$ with $f \in \mathcal{F}$ has a single basic class).

If Assumption B holds we can show (e.g. by policy iterations) existence of numbers w_i^* $(i \in \mathcal{I})$, g^*, and some $f^* \in \mathcal{F}$ such that for all $i \in \mathcal{I}$

$$\sum_{j \in \mathcal{I}} p_{ij}(f_i) e^{\gamma\{r_{ij} + w_j^*\}} \leq \sum_{j \in \mathcal{I}} p_{ij}(f_i^*) e^{\gamma\{r_{ij} + w_j^*\}} = \mathrm{e}^{\gamma[g^* + w_i^*]}, \tag{34}$$

or equivalently

$$\sum_{j \in \mathcal{I}} q_{ij}(f_i) z_j(f^*) \leq \sum_{j \in \mathcal{I}} q_{ij}(f_i^*) z_j(f^*) = \rho(f^*) z_i(f^*). \tag{35}$$

Similarly if Assumption B is fulfilled there also exist $\hat{w}_i$ $(i \in \mathcal{I})$, $\hat{g}$, and some $\hat{f} \in \mathcal{F}$

such that for all $i \in \mathcal{I}$

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)e^{\gamma\{r_{ij}+\hat{w}_j\}} \geq \sum_{j\in\mathcal{I}} p_{ij}(\hat{f}_i)e^{\gamma\{r_{ij}+\hat{w}_j\}} = \mathrm{e}^{\gamma[\hat{g}+\hat{w}_i]}, \tag{36}$$

or equivalently

$$\sum_{j\in\mathcal{I}} q_{ij}(f_i)z_j(\hat{f}) \geq \sum_{j\in\mathcal{I}} q_{ij}(\hat{f}_i)z_j(\hat{f}) = \rho(\hat{f})z_i(\hat{f}). \tag{37}$$

Observe that by (35), (37) it holds for any $f \in \mathcal{F}$

$$Q(f)z(f^*) \leq Q(f^*)z(f^*) = \rho(f^*)z(f^*), \quad Q(f)z(\hat{f}) \geq Q(\hat{f})z(\hat{f}) = \rho(\hat{f})z(\hat{f}). \tag{38}$$

The above facts can be summarized in the following theorem.

Theorem 3. *If Assumption B holds, there exists decision vector $f^* \in \mathcal{F}$ (resp. $\hat{f} \in \mathcal{F}$) along with strictly positive column vector $z(f^*)$ (resp. $z(\hat{f})$) and a positive number $\rho(f^*)$, along with $g(f^*) = \ln\rho(f^*)$, (resp. $\rho(\hat{f})$, along with $g(\hat{f}) = \ln\rho(\hat{f})$) such that (34)–(38) hold and for any $f \in \mathcal{F}$ $\rho(\hat{f}) \leq \rho(f) \leq \rho(f^*)$, $g(\hat{f}) \leq g(f) \leq g(f^*)$.*

The proof (by policy iterations) based on ideas in Howard and Matheson (1972) can be found in Sladký (2008b).

6. Necessary and sufficient optimality conditions

6.1 Risk-neutral case

To begin with, from Eq. (19) considered for decision vector f^* maximizing the average reward with $g = g^*$, $w = w^*$, we immediately have for policy $\pi = (f^n)$

$$V(\pi,n) := \sum_{k=0}^{n-1}\prod_{j=0}^{k-1} P(f^j)r(f^k) = ng^* + w^* + \sum_{k=0}^{n-1}\prod_{j=0}^{k-1} P(f^j)\varphi(f^k,f^*) - \prod_{j=0}^{n} P(f^j)w^*. \tag{39}$$

Hence for stationary policy $\pi^* \sim (f^*)$ maximizing average reward, we immediately get

$$V(\pi^*,n) = ng^* + w^* - \prod_{j=0}^{n} P(f^j)w^* \tag{40}$$

and (cf. Mandl 1971; Sladký 1974) nonstationary policy $\pi = (f^n)$ maximizes long run average reward if and only if

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1}\prod_{j=0}^{k-1} P(f^j)\varphi(f^k,f^*) = 0. \tag{41}$$

6.2 Risk-sensitive case

From Eq. (31) considered for decision vector f^* fulfilling conditions (34), (35), we immediately have for policy $\pi = (f^n)$

$$U_i^\gamma(\pi,n) = \mathrm{e}^{\gamma(g^*+w_i^*)} \sum_{j\in\mathcal{I}} p_{ij}(f_i^0) e^{\gamma\{\overline{\varphi}_{ij}(w^*,g^*)-w_j^*\}} \cdot U_j^\gamma(\pi^1,n-1) \tag{42}$$

$$= \mathrm{e}^{\gamma(ng^*+w_i^*)} \mathrm{E}_i^\pi e^{\gamma\{\sum_{k=0}^{n-1}\overline{\varphi}_{X_k,X_{k+1}}(w^*,g^*)-w_{X_n}^*\}}, \tag{43}$$

and for stationary policy $\pi^* \sim (f^*)$ with f^* fulfilling conditions (36), (37), we have

$$U_i^\gamma(f^*,n) = \mathrm{e}^{\gamma(ng^*+w_i^*)} \mathrm{E}_i^\pi e^{\{-\gamma w_{X_n}^*\}}. \tag{44}$$

Similarly,

$$U_i^\gamma(\pi,n) = \mathrm{e}^{\gamma(n\hat{g}+\hat{w}_i)} \mathrm{E}_i^\pi e^{\gamma\{\sum_{k=0}^{n-1}\overline{\varphi}_{X_k,X_{k+1}}(\hat{w},\hat{g})-\hat{w}_{X_n}\}}. \tag{45}$$

Recalling that $Z_i^\gamma(\pi,n) = \frac{1}{\gamma}\ln U_i^\gamma(\pi,n)$, let

$$g_i'(\pi) := \liminf_{n\to\infty} \frac{1}{n} Z_i^\gamma(\pi,n), \quad g_i''(\pi) := \limsup_{n\to\infty} \frac{1}{n} Z_i^\gamma(\pi,n)$$

be the corresponding mean values of $Z_i^\gamma(\pi,n)$.

Theorem 4. *Let Assumption B hold and $g^* = \ln\rho(f^*)$, $\hat{g} = \ln\rho(\hat{f})$. Then*

$$\lim_{n\to\infty} \frac{1}{n} Z_i^\gamma(\pi,n) = g^* \quad \textit{if and only if} \quad \lim_{n\to\infty} \frac{1}{n} \ln\left\{E_i^\pi \mathrm{e}^{\gamma\sum_{k=0}^{n-1}\overline{\varphi}_{X_k,X_{k+1}}(w^*,g^*)}\right\} = 0, \tag{46}$$

$$\lim_{n\to\infty} \frac{1}{n} Z_i^\gamma(\pi,n) = \hat{g} \quad \textit{if and only if} \quad \lim_{n\to\infty} \frac{1}{n} \ln\left\{E_i^\pi \mathrm{e}^{\gamma\sum_{k=0}^{n-1}\overline{\varphi}_{X_k,X_{k+1}}(\hat{w},\hat{g})}\right\} = 0. \tag{47}$$

Proof. Since the state space $\mathcal{I}$ is finite, there exists number $K > 0$ such that $|w_i^*| \le K$ for each $i \in \mathcal{I}$. Hence by (43), (45) we immediately conclude that

$$\mathrm{e}^{\gamma(n\hat{g}+\hat{w}_i)} \cdot \mathrm{e}^{-|\gamma|K} \le U_i^\gamma(\pi,n) \le \mathrm{e}^{\gamma(ng^*+w_i^*)} \cdot \mathrm{e}^{|\gamma|K},$$

$$n\hat{g} + \hat{w}_i + const. \le Z_i^\gamma(\pi,n) = \frac{1}{\gamma}\ln U_i^\gamma(\pi,n) \le ng^* + w_i^* + const.$$

and (46), (47) follow by (43), (45). □

7. Conclusions

In this note we focused attention on necessary and sufficient optimality conditions for more sophisticated optimality criteria in unichain Markov decision processes with finite state space taking into account also the variability risk features of the model. To this end, necessary and sufficient mean reward optimality conditions for unichain

Markov models were obtained, and using the formulas for mean variance we can select in the class of mean optimal control the policy minimizing the mean variance. Another approach for the so-called risk-sensitive optimality is based on replacing linear utility function by exponential utilities that are also separable and hence suitable for sequential decision. Using some results of a family of nonnegative matrices necessary and sufficient optimality conditions for risk-sensitive optimality are obtained if the underlying Markov process is either irreducible, or unichain with the risk sensitive coefficient sufficiently close to zero.

Acknowledgement This paper is an extended version of the author's paper entitled "Risk sensitive and risk-neutral optimality in Markov decision chains; a unified approach" presented at the International Scientific Conference Quantitative Methods in Economics (Multiple Criteria Decision Making XVI) held in Bratislava (Slovak Republic) on May 30–June 1, 2012. This research was supported by the Czech Science Foundation under Grants P402/11/0150 and P402/10/0956, and by CONACyT (México) and ASCR (Czech Republic) under Project 171396. The author is indebted to two anonymous referees for insightful reading the manuscript and helpful comments.

References

Borkar, V. and Meyn, S. P. (2002). Risk-Sensitive Optimal Control for Markov Decision Process with Monotone Costs. *Mathematics of Operations Research,* 27, 192–209.

Cavazos-Cadena, R. (2002). Value Iteration and Approximately Optimal Stationary Policies in Finite-State Average Markov Chains. *Mathematical Methods of Operations Research,* 56, 181–196.

Cavazos-Cadena, R. (2003). Solution to the Risk-Sensitive Average Cost Optimality Equation in a Class of Markov Decision Processes with Finite State Space. *Mathematical Methods of Operations Research,* 57, 253–285.

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, 121–139.

Cavazos-Cadena, R. and Fernandez-Gaucherand, F. (2000). The Vanishing Discount Approach in Markov Chains with Risk-Sensitive Criteria. *IEEE Transactions on Automatic Control,* 45, 1800–1816.

Cavazos-Cadena, R. and Hernández-Hernández, D. (2002). Solution of the Risk-Sensitive Average Optimality Equation in Communicating Markov Decision Chains with Finite State Space: An Alternative Approach. *Mathematical Methods of Operations Research,* 56, 473–479.

Cavazos-Cadena, R. and Hernández-Hernández, D. (2004). A Characterization Exponential Functionals in Finite Markov Chains. *Mathematical Methods of Operations Research,* 60, 399–414.

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, 175–212.

Cavazos-Cadena, R. and Montes-de-Oca, R. (2003). The Value Iteration Algorithm in Risk-Sensitive Average Markov Decision Chains with Finite State Space. *Mathematics of Operations Research,* 28, 752–756.

Cavazos-Cadena, R. and Montes-de-Oca, R. (2005). Nonstationary Value Iteration in Controlled Markov Chains with Risk-Sensitive Average Criterion. *Journal of Applied Probability,* 42, 905–918.

Di Masi, G. B. and Stettner, L. (1999). Risk-Sensitive Control of Discrete Time Markov Processes with Infinite Horizon. *SIAM Journal on Control and Optimization,* 38, 61–78.

Di Masi, G. B. and Stettner, L. (2000). Infinite Horizon Risk Sensitive Control of Discrete Time Markov Processes with Small Risk. *Systems and Control Letters,* 40, 15–20.

Di Masi, G. B. and Stettner, L. (2006). On Additive and Multiplicative (Controlled) Poisson Equations: Approximation and Probability. *Banach Center Publications,* 72, 57–60.

Di Masi, G. B. and Stettner, L. (2007). Infinite Horizon Risk Sensitive Control of Discrete Time Markov Processes under Minorization Property. *SIAM Journal on Control and Optimization,* 46, 231–252.

Filar, J., Kallenberg, L. C. M. and Lee, H.-M. (1989). Variance Penalized Markov Decision Processes. *Mathematics of Operations Research,* 14, 147–161.

Gantmakher, F. R. (1959). *The Theory of Matrices.* London, Chelsea Publishing Series.

Hernández-Hernández, D. and Marcus, S. I. (1999). Existence of Risk Sensitive Optimal Stationary Policies for Controlled Markov Processes. *Applied Mathematics and Optimization,* 40, 273–285.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes.* Cambridge MA, MIT Press.

Howard, R. A. and Matheson, J. (1972). Risk-Sensitive Markov Decision Processes. *Management Science,* 23, 356–369.

Huang, Y. and Kallenberg, L. C. M. (1994). On Finding Optimal Policies for Markov Decision Chains: A Unifying Framework for Mean-Variance-Tradeoffs.*Mathematics of Operations Research,* 19, 434–448.

Jaquette, S. A. (1976). A Utility Criterion for Markov Decision Processes. *Management Science,* 23, 43–49.

Kadota, Y. (1997). A Minimum Average-Variance in Markov Decision Processes. *Bulletin of Informatics and Cybernetics,* 29, 83–89.

Kawai, H. (1987). A Variance Minimization Problem for a Markov Decision Process. *European Journal of Operational Research,* 31, 140–145.

Kurano, M. (1987). Markov Decision Processes with a Minimum-Variance Criterion. *Journal of Mathematical Analysis and Applications,* 123, 572–583.

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

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

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

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

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

Sladký, K. (1974). On the Set of Optimal Controls for Markov Chains with Rewards. *Kybernetika,* 10, 526–547.

Sladký, K. and Sitař, M. (2004). Optimal Solutions for Undiscounted Variance Penalized Markov Decision Chains. In Marti, K., Ermoliev, Y. and Pflug, G. (eds.), *Dynamic Stochastic Optimization, Lecture Notes in Economics and Mathematical Systems, Vol. 532*. Berlin-Heidelberg, Springer-Verlag, 43–66.

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

Sladký, K. (2008a). Risk Sensitive Discrete- and Continuous-Time Markov Reward Processes. In Reiff, M. (ed.), *Proceedings of the International Scientific Conference Quantitative Methods in Economics (Multiple Criteria Decision Making XIV)*. Bratislava, University of Economics, 272–281.

Sladký, K. (2008b). Growth Rates and Average Optimality in Risk-Sensitive Markov Decision Chains. *Kybernetika*, 44, 205–226.

Sladký, K. (2012). Risk Sensitive and Risk-Neutral Optimality in Markov Decision Chains: A Unified Approach. In Reiff, M. (ed.), *Proceedings of the International Scientific Conference Quantitative Methods in Economics (Multiple Criteria Decision Making XIV)*. Bratislava, University of Economics, 201–205.

Sobel, M. J. (1985). Maximal Mean/Standard Deviation Ratio in an Undiscounted MDP. *Operations Research Letters,* 4, 157–159.

Appendix

Proof of Theorem 2. (Based on Sladky 2005.)

For the sake of simplicity in what follows we omit arguments π, f, e.g. E_i, $V_i(n)$, $S_i(n)$, $\sigma_i(n)$ respectively denotes the conditional expectation, expected reward up to n, its second moment and variance respectively if policy $\pi \sim (f)$ is followed and $X(0) = i$, p_{ij} is the transition probability if action f_i is chosen in state $i \in \mathcal{I}$.

For the variance $\sigma_i(\cdot) = S_i(\cdot) - [V_i(\cdot)]^2$ we have by (15)

$$\begin{aligned} \sigma_i(n+1) &= r_i^{(2)} + \sum_{j\in\mathcal{I}} p_{ij}\cdot\sigma_j(n) + 2\cdot\sum_{j\in\mathcal{I}} p_{ij}\, r_{ij}\cdot V_j(n) \\ &\quad - \sum_{j\in\mathcal{I}} p_{ij}\cdot[V_i(n+1)+V_j(n)]\cdot[V_i(n+1)-V_j(n)]. \end{aligned} \tag{A1}$$

If P is aperiodic by (24) in the last term of (A1), we can substitute:

$$V_i(n+1)+V_j(n) = 2\cdot n\cdot g + g + w_i + w_j + \varepsilon(n), \tag{A2}$$

$$V_i(n+1)-V_j(n) = g + w_i - w_j + \varepsilon(n), \tag{A3}$$

where $\varepsilon(n) \to 0$ geometrically. Hence by (22)

$$\begin{aligned} &\sum_{j\in\mathcal{I}} p_{ij}\cdot[V_i(n+1)+V_j(n)]\cdot[V_i(n+1)-V_j(n)] = \\ &= 2\cdot n\cdot g\cdot(g+w_i-\sum_{j\in\mathcal{I}} p_{ij}\cdot w_j) + \sum_{j\in\mathcal{I}} p_{ij}\cdot\{[g+w_i]^2-[w_j]^2\}+\varepsilon(n) \\ &= 2\cdot n\cdot g\cdot\ r_i^{(1)} + \sum_{j\in\mathcal{I}} p_{ij}\cdot\{[g+w_i]^2-[w_j]^2\}+\varepsilon(n), \end{aligned} \tag{A4}$$

where $\lim_{n\to\infty}\varepsilon(n) = 0$ and the convergence is geometrical (cf. (24),(25)).

Similarly for the third term of (A1) we obtain by (24)

$$\begin{aligned} \sum_{j\in\mathcal{I}} p_{ij}\, r_{ij}\, V_j(n) &= \sum_{j\in\mathcal{I}} p_{ij}\, r_{ij}\,[n\cdot g + w_j + \varepsilon(n)] \\ &= n\cdot g\cdot\ r_i^{(1)} + \sum_{j\in\mathcal{I}} p_{ij}\, r_{ij}\cdot w_j + \varepsilon(n). \end{aligned} \tag{A5}$$

Substituting (A5) and (A4) in (A1) now yields

$$\begin{aligned}\sigma_i(n+1) &= \sum_{j\in\mathcal{I}} p_{ij}\cdot\sigma_j(n) + r_i^{(2)} + 2\cdot\sum_{j\in\mathcal{I}} p_{ij}\, r_{ij}w_j \\ &\quad - \sum_{j\in\mathcal{I}} p_{ij}\cdot\{[g+w_i]^2 - [w_j]^2\} + \varepsilon(n) \\ &= \sum_{j\in\mathcal{I}} p_{ij}\cdot\sigma_j(n) + s_i + \varepsilon(n), \qquad \text{(A6)}\end{aligned}$$

where

$$\begin{aligned} s_i &= r_i^{(2)} + \sum_{j\in\mathcal{I}} p_{ij}\big\{[w_j]^2 + 2\, r_{ij}w_j\big\} - [g+w_i]^2 \\ &= \sum_{j\in\mathcal{I}} p_{ij}[\, r_{ij}+w_j]^2 - [g+w_i]^2 \qquad \text{(A7)} \\ &= \sum_{j\in\mathcal{I}} p_{ij}[\, r_{ij}+w_j-g]^2 - [w_i]^2\ . \qquad \text{(A8)}\end{aligned}$$

(A8) follows immediately since

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