Risk-Sensitive Optimality in Markov Games

Karel Sladký[1], Victor Manuel Martínez Cortés[2]

Abstract. The article is devoted to risk-sensitive optimality in Markov games. Attention is focused on Markov games evolving on communicating Markov chains with two-players with opposite aims. Considering risk-sensitive optimality criteria means that total reward generated by the game is evaluated by exponential utility function with a given risk-sensitive coefficient. In particular, the first player (resp. the second player) tries to maximize (resp. minimize) the long-run risk-sensitive average reward. Observe that if the second player is dummy, the problem is reduced to finding optimal policy of the Markov decision chain with the risk-sensitive optimality. Recall that for the risk sensitivity coefficient equal to zero we arrive at traditional optimality criteria. In this article, connections between risk-sensitive and risk-neutral Markov decision chains and Markov games models are studied using discrepancy functions. Explicit formulae for bounds on the risk-sensitive average long-run reward are reported. Policy iteration algorithm for finding suboptimal policies of both players is suggested. The obtained results are illustrated on numerical example.

Keywords: dynamic programming, Markov decision chains, two-person Markov games, communicating Markov chains, risk-sensitive optimality.

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

1 Introduction

This contribution is devoted to risk-sensitive optimality in Markov games evolving on communicating Markov chains with two-players with opposite aims. In particular, the first player (resp. the second player) tries to maximize (resp. minimize) the long-run risk-sensitive average reward calculated by an exponential utility function with a given risk-sensitive coefficient. Observe that if the second player is dummy, the problem is reduced to finding optimal policy of the risk-sensitive Markov decision chain introduced by Howard and Matheson in their seminal paper [8]. Recall that for the risk sensitivity coefficient equal to zero we arrive at traditional optimality criteria. In this article, connections between risk-sensitive and risk-neutral Markov decision chains and Markov games models are studied using discrepancy functions. Explicit formulae for bounds on the risk-sensitive average long-run reward are reported. Policy iteration algorithms for finding suboptimal policies of both players are suggested.

2 Notation and Preliminaries

In this note, we consider at discrete time points $t = 0, 1, \ldots$ a dynamic system $X = \{X_n, n = 0, 1, \ldots\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$. The behavior of the system X is influenced by two players, $P^{(1)}$ and $P^{(2)}$, with opposite aims. Supposing that at time t the system is in state $i \in \mathcal{I}$ then player $P^{(1)}$, resp. player $P^{(2)}$, selects action $a^{(1)}$ from finite set $\mathcal{A}_i^{(1)}$, resp. action $a^{(2)}$ from finite set $\mathcal{A}_i^{(2)}$. Then state j is reached in the next transition with a given probability $p_{ij}(a^{(1)}, a^{(2)})$ and one-stage reward $r_i(a^{(1)}, a^{(2)})$ is accrued. We shall call this two person game a Markov game.

In this note, we assume that the stream of rewards generated by the Markov processes X is evaluated by an exponential utility function (so-called risk-sensitive models) with a given risk sensitivity coefficient. To this end, let us consider an exponential utility function, say $\bar{u}^{\gamma}(\cdot)$, i.e. a separable utility function with constant risk sensitivity $\gamma \in \mathbb{R}$. Then the utility assigned to the (random) outcome ξ is given by

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

Obviously $\bar{u}^{\gamma}(\cdot)$ is continuous and strictly increasing. For $\gamma > 0$ $\bar{u}^{\gamma}(\cdot)$ is convex, if $\gamma < 0$ $\bar{u}^{\gamma}(\cdot)$ is concave. Finally if $\gamma = 0$ (risk neutral case) $\bar{u}^{\gamma}(\cdot)$ is linear. Observe that exponential utility function $\bar{u}^{\gamma}(\cdot)$ is separable and

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

[2]Department of Mathematics, Autonomous Metropolitan University, Iztapalapa Campus, Mexico, mat.victor.m.mtz.c@gmail.com

multiplicative if the risk sensitivity $\gamma \neq 0$ and additive for $\gamma = 0$. In particular, for $u^\gamma(\cdot) := \exp(\gamma\xi)$ 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 $\gamma = 0$.

Moreover, recall that 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

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

The development of the system X over time is controlled by actions of both players that have complete information about the history of the system. In particular, player $P^{(1)}$, resp. player $P^{(2)}$, tries to maximize, resp. minimizes the total reward. Supposing that the system is in state $i \in \mathcal{I}$ if decision $a^{(1)} \in \mathcal{A}_i^{(1)}$ is taken by the first player, player $P^{(2)}$ selects decision $a^{(2)} \in \mathcal{A}_i^{(2)}$ such that to "minimize" possible outcome (so decisions $a^{(1)}, a^{(2)}$ are not simultaneous, player $P^{(1)}$ is the leader and player $P^{(2)}$ is the follower in the considered Stackelberg duopoly model). Risk-sensitive Markov decision chains can be considered as a special case of Markov games with only one player.

A (Markovian) policy controlling the decision process, $\pi = (f^0, f^1, \ldots)$, is identified by a sequence of decision vectors $\{f^n, n = 0, 1, \ldots\}$ where $f^n = (f^{(1)n}, f^{(2)n}) \in \mathcal{F} \equiv \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$. In particular, player $P^{(1)}$, resp. player $P^{(2)}$, generates a sequence of decisions $f^{(1),n}$ where $f^{(1),n} \in \mathcal{F}^{(1)} \equiv \mathcal{A}_1^{(1)} \times \ldots \times \mathcal{A}_N^{(1)}$, resp. $f^{(2),n}$ where $f^{(2),n} \in \mathcal{F}^{(2)} \equiv \mathcal{A}_1^{(2)} \times \ldots \times \mathcal{A}_N^{(2)}$.

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, $\mathsf{E}_i^\pi(X_m = j) = \sum_{i_j \in \mathcal{I}} p_{i,i_1}(f_i^0) \ldots p_{i_{m-1},j}(f_{m-1}^{m-1})$; $\mathrm{P}(X_m = j)$ 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 following policy $\pi \sim (f)$ X is a homogeneous Markov chain with transition probability matrix $P(f)$ whose ij-th element is $p_{ij}(f_i) = p_{ij}(f_i^{(1)}, f_i^{(2)})$. Then $r_i(f_i) := r_i(f_i^{(1)}, f_i^{(2)})$ is the one-stage reward obtained in state i. Similarly, $r(f)$ is an N-column vector of one-stage rewards whose i-the elements equals $r_i(f_i)$.

Stationary policy $\tilde{\pi}$ is randomized if there exist decision vectors $f^{[1]}, f^{[2]}, \ldots, f^{[m]} \in \mathcal{F}$ (observe that $f^{[1]} = (f^{[1](1)}, f^{[1](2)}) \in \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$). 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 $i \in \mathcal{I}$). Observe that $\mathsf{E}_i^\pi(X_m = j) = [P^m(f)]_{ij}$ (here $[A]_{ij}$ denotes the ij-th element of the matrix A, $A \geq B$, resp. $A > B$ iff for each i, j $[A]_{ij} \geq [B]_{ij}$ resp. $[A]_{ij} \geq [B]_{ij}$ and $[A]_{ij} > [B]_{ij}$ for some i, j). The symbol I denotes an identity matrix and e is reserved for a unit column vector.

3 Risk-Sensitive Optimality in Markov Processes

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

On introducing for arbitrary $g, w_j \in \mathbb{R}$ $(i, j \in \mathcal{I})$ the discrepancy function (cf. [10]) $\tilde{\varphi}_{i,j}(w, g) := r_i - w_i + w_j - g$ we can easily verify the following identity:

$$\xi_n \quad = \quad ng + w_{X_0} - w_{X_n} + \sum_{k=0}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w, g). \tag{4}$$

Considering the risk-sensitive models in virtue of (1), (4) for the expectation of ξ_n in the risk-sensitive case $U_i^\pi(\gamma, n) := \mathsf{E}_i^\pi \mathrm{e}^{\gamma \sum\limits_{k=0}^{n-1} \xi_n}$ we conclude that

$$U_i^\pi(\gamma, n) = \mathrm{e}^{\gamma[ng + w_i]} \times \mathsf{E}_i^\pi \mathrm{e}^{\gamma[\sum\limits_{k=0}^{n-1} \tilde{\varphi}_{X_k, X_{k+1}}(w,g) - w_{X_n}]}. \tag{5}$$

Now observe that

$$\mathsf{E}_i^{\pi}\, e^{\gamma \sum_{k=0}^{n-1} \tilde\varphi_{X_k,X_{k+1}}(w,g)} = \sum_{j\in\mathcal{I}} p_{ij}(f_i^0)\, e^{\gamma[r_i - w_i + w_j - g]} \times \mathsf{E}_j^{\pi^1}\, e^{\gamma \sum_{k=1}^{n-1} \tilde\varphi_{X_k,X_{k+1}}(w,g)} \tag{6}$$

$$\mathsf{E}_j^{\pi}\, \{e^{\gamma \sum_{k=m}^{n-1} \tilde\varphi_{X_k,X_{k+1}}(w,g)} | X_m = j\} = \sum_{\ell\in\mathcal{I}} p_{j,\ell}(f_j^m)\, e^{\gamma[r_j - w_j + w_\ell - g]} \times \mathsf{E}_\ell^{\pi^{m+1}}\, e^{\gamma \sum_{k=m+1}^{n-1} \tilde\varphi_{X_k,X_{k+1}}(w,g)}. \tag{7}$$

If stationary policy $\pi \sim (f)$ is followed (5) can be drastically simplified if the numbers g, w_j's are selected such that $\sum_{j\in\mathcal{I}} p_{ij}(f_i)\, e^{\gamma\tilde\varphi_{ij}(g,w)} = 1$ for all $i \in \mathcal{I}$.[3] Obviously, this condition is equivalent to the following set of linear equations

$$e^{\gamma[g(f)+w_i(f)]} = \sum_{j\in\mathcal{I}} p_{ij}(f_i)\; e^{\gamma[r_i(f_i)+w_j(f)]} \quad (i \in \mathcal{I}) \tag{8}$$

for the values $g(f), w_i(f) (i = 1,\dots,N)$; observe that these values depend on the selected risk sensitivity γ. Eqs. (5) can be called the *γ-average reward/cost optimality equation.*

On introducing the new variables $v_i(f) := e^{\gamma w_i(f)}$, $\rho(f) := e^{\gamma g(f)}$, and on replacing transition probabilities $p_{ij}(f_i)$'s by general nonnegative numbers defined by $q_{ij}(f_i) := p_{ij}(f_i) \cdot e^{\gamma r_i(f_i)}$ (8) can be alternatively written as the following set of equations

$$\rho(f) v_i(f) = \sum_{j\in\mathcal{I}} q_{ij}(f_i)\, v_j(f) \quad (i \in \mathcal{I}) \tag{9}$$

For what follows it is convenient to consider (9) in matrix form. To this end, we introduce (cf. [6]) $N \times N$ matrix $Q(f) = [q_{ij}(f_i)]$ with spectral radius (Perron eigenvalue) $\rho(f)$ along with its right Perron eigenvector $v(f) = [v_i(f_i)]$. Then (9) can be written in matrix form as

$$\rho(f) v(f) = Q(f) v(f). \tag{10}$$

Furthermore, if the transition probability matrix $P(f)$ is *irreducible* then also $Q(f)$ is *irreducible* and the right Perron eigenvector $v(f)$ can be selected *strictly positive.*

From (3),(5),(6),(8) we immediately get for stationary policy $\pi \sim (f)$ that

$$U_i^{\pi}(\gamma, n) = e^{\gamma[ng(f)+w_i(f)]} \times \mathsf{E}_i^{\pi} e^{\gamma w_{X_n}(f)}, \quad Z_i^{\pi}(\gamma, n) = \frac{1}{\gamma} \ln U_i^{\pi}(\gamma, n).$$

Hence

$$n^{-1} Z_i^{\pi}(\gamma, n) = g(f) + o(n) \tag{11}$$

(recall that $g(f) = \gamma^{-1} \ln \rho(f)$, $w_i(f) = \gamma^{-1} \ln v_i(f)$).

If the Markov chain is irreducible there exist $\hat f, f^* \in \mathcal{F}$ along with numbers $\hat\rho = \rho(\hat f)$, $\rho^* = \rho(f^*)$ and strictly positive vectors $\hat v = v(\hat f)$, with elements $v_i(\hat f)$ and $v^* = v(f^*)$ with elements $v_i(f^*)$ such that for any $f \in \mathcal{F}$ (vectorial max and min should be considered componentwise)

$$Q(f) \cdot \hat v \ge \min_{f\in\mathcal{F}} \{Q(f) \cdot \hat v\} = Q(\hat f) \cdot \hat v = \hat\rho \cdot \hat v \tag{12}$$

$$Q(f) \cdot v^* \le \max_{f\in\mathcal{F}} \{Q(f) \cdot v^*\} = Q(f^*) \cdot v^* = \rho^* \cdot v^* \tag{13}$$

$$\rho(\hat f) \equiv \hat\rho \le \rho(f) \le \rho(f^*) \equiv \rho^* \qquad \text{for all } f \in \mathcal{F}. \tag{14}$$

In words:

$\hat\rho \equiv \rho(\hat f)$ (resp. $\rho^* = \rho(f^*)$) is the minimum (resp. maximum) possible eigenvalue of $Q(f)$ over all $f \in \mathcal{F}$ (cf. [1],[3],[8]).

Minimal (resp. maximal) risk-sensitive average reward $g(\hat f) = \gamma^{-1} \ln \rho(\hat f)$ (resp. $g(f^*) = \gamma^{-1} \ln \rho(f^*)$).

[3]To verify this claim it suffices to apply successively (7) backwards starting time point $n - 1$ (cf. [8]).

4 Risk-Sensitive Optimality in Markov Games

In contrast to Markov decision model considered in section 3 we assume that the expected utility $U_i^\pi(\gamma, n)$ depends on decision $f^{(1),n}$, $f^{(2),n}$ taken by the both players. Since Markov decision processes can be considered as a very special case of Markov games, it is interesting to mention that stochastic games were formulated by Shapley [12] in 1953, many years before outburst of systematic interest in Markov decision processes. For the early results on Markov decision processes see Bellman's papers [1], [2], Bellman's monograph [3], Blackwell's paper [4] and especially Howard's book [7].

In contrast to Markov decision processes we must take into consideration decision taken by both players. Hence the optimality equations (12), (13) must be replaced by the Nash equilibrium condition, see [11]. According to the Nash equilibrium condition there exist $f^* = (f^{(1)*}, f^{(2)*}) \in \mathcal{F} = \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$ such that for any $f_i^{(1)} \in \mathcal{F}_i^{(1)}$ and any $f_i^{(2)} \in \mathcal{F}_i^{(2)}$ for the resulting decisions $f_i^d = (f_i^{(1)}, f_i^{(2)*})$, $f_i^u = (f_i^{(1)*}, f_i^{(2)})$, it holds

$$\sum_{j\in\mathcal{I}} q_{ij}(f_i^d)v_j^* \le \sum_{j\in\mathcal{I}} q_{ij}(f_i^*)v_j^* = \rho(f^*)v_i^* \le \sum_{j\in\mathcal{I}} q_{ij}(f_i^u)v_j^*, \tag{15}$$

or in matrix notations $\rho(f)v(f) = Q(f)v(f)$ we are looking for $f^* = (f^{(1)*}, f^{(2)*}) \in \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$ such that

$$Q(f^d)v(f^*) \le \rho(f^*)v(f^*) = Q(f^*)v(f^*) \le Q(f^u)v(f^*) \tag{16}$$

where $\rho(f^*) = \mathrm{e}^{\gamma g(f^*)}$, $v_i(f^*) = \mathrm{e}^{\gamma w_i(f^*)}$.

From (3),(5),(6),(8) we immediately get for stationary policy $\pi \sim (f)$ that

$$U_i^{\pi^*}(\gamma, n) = \mathrm{e}^{\gamma[ng(f^*)+w_i(f^*)]} \times \mathsf{E}_i^{\pi^*} \mathrm{e}^{\gamma w_{X_n}(f^*)}, \quad Z_i^{\pi^*}(\gamma, n) = \frac{1}{\gamma} \ln U_i^{\pi^*}(\gamma, n), \tag{17}$$

$$n^{-1} Z_i^{\pi^*}(\gamma, n) = g(f^*) + o(n). \tag{18}$$

Since the average risk-sensitive reward $g(f) = \gamma^{-1}\ln[\rho(f)]$ and $\rho(f)$ is the Perron eigenvalue of a nonnegative matrix $Q(f)$, it is well-known (see e.g. [6]) that for any $f^{'}, f^{''} \in \mathcal{F}$ $Q(f^{'}) \le Q(f^{''}) \Rightarrow \rho(f^{'}) \le \rho(f^{''})$. To generate lower and upper bounds on minimal and maximal Perron eigenvalue $\rho(f^*)$ we replace elements $q_{ij}(f_i^{(1)}, f_i^{(2)})$ by their minimal and maximal possible values $q_{ij}^{'}$ and $q_{ij}^{''}$. Then the problem is approximated by a (uncontrollable) risk-sensitive Markov chain and it is possible to generate lower and upper bounds on $\rho(f^*) = \mathrm{e}^{\gamma g(f^*)}$ by calculating Perron eigenvalues (i.e. the spectral radii) of nonnegative matrices. Unfortunately, using this approach we can expect only very rough bounds on the optimal value of the average risk-sensitive reward.

More friendly bounds can be obtained by a more detailed analysis of the set of all admissible matrices. Of course, it is reasonable to suggest algorithmic procedures that need not evaluate all admissible matrices. Algorithm 1 is a slight modification of the policy iteration method reported in [8] only for finding maximum Perron eigenvalue in a set of nonnegative irreducible matrices.

Algorithm 1. (Policy iterations for finding maximal, resp. minimal, Perron eigenvalue.)

Step 0. Find matrix $Q^{(0)} := Q(f^{(1),0}, f^{(2),0})$ with $f^{(1),0} \in \mathcal{F}^{(1)}$, $f^{(2),0} \in \mathcal{F}^{(2)}$ such that the row sums are maximal (resp. minimal).
Step 1. For matrix $Q^{(k)}$ $(k = 0, 1, \ldots)$ calculate its spectral radius $\rho^{(k)}$ along with its right Perron eigenvector $v^{(k)}$.
Step 2. Construct (if possible) matrix $Q^{(k+1)} := Q(f^{(1),k+1}, f^{(2),k+1})$ with
$f^{k+1} := (f^{(1),k+1}, f^{(2),k+1})$ where $f^{(1),k+1} \in \mathcal{F}^{(1)}$, $f^{(2),k+1} \in \mathcal{F}^{(2)}$, such that

$$Q^{(k+1)} \cdot v^{(k)} \ge \rho^{(k)}\, v^{(k)} = Q^{(k)} \cdot v^{(k)} \quad \text{resp.} \quad Q^{(k+1)} \cdot v^{(k)} \le \rho^{(k)}\, v^{(k)} = Q^{(k)} \cdot v^{(k)} \tag{19}$$

Step 3. If $Q^{(k+1)} = Q^{(k)}$ then go to Step 4, else set $k := k + 1$ and repeat Step 1.
Step 4. Set $\hat{Q} := Q^{(k+1)}$, $\hat{\rho} := \rho^{(k+1)}$, $\hat{v} := v^{(k+1)}$, $\hat{f} := f^{(k+1)}$ and stop. $\hat{\rho}$ is the maximal (resp. minimal) Perron eigenvalue.

The heart of the above algorithms is the following

Policy improvement routine:
Since for the right (resp. left) Perron eigenvectors $v^{(k)}$ (resp. $z^{(k)}$) of an irreducible matrix $Q^{(k)}$ it holds $Q^{(k)} \cdot v^{(k)} = \rho^{(k)} v^{(k)}$ (resp. $z^{(k)} Q^{(k)} = \rho^{(k)} z^{(k)}$) if $\varphi^{(k+1)} := Q^{(k+1)} \cdot v^{(k)} - Q^{(k)} \cdot v^{(k)} > 0$ (resp. < 0) then

$$Q^{(k+1)} \cdot v^{(k+1)} - Q^{(k)} \cdot v^{(k)} = \rho^{(k+1)}[v^{(k+1)} - v^{(k)}] + [\rho^{(k+1)} - \rho^{(k)}]v^{(k)}$$

On premultiplying the above equality by $z^{(k+1)}$ (strictly positive row vector) we arrive at

$$\rho^{(k+1)} \cdot z^{(k+1)}[v^{(k+1)} - v^{(k)}] + [\rho^{(k+1)} - \rho^{(k)}] \cdot z^{(k+1)} v^{(k)} = z^{(k+1)} Q^{(k+1)}[v^{(k+1)} - v^{(k)}] + z^{(k+1)} \varphi^{(k+1)}$$

implying that $z^{(k+1)}\varphi^{(k+1)} = [\rho^{(k+1)} - \rho^{(k)}] z^{(k+1)} v^{(k)}$.
Since $z^{(k+1)} v^{(k)} > 0$ if $z^{(k+1)}\varphi^{(k+1)} > 0$ (resp. $z^{(k+1)}\varphi^{(k+1)} < 0$) then $\rho^{(k+1)} > \rho^{(k)}$ (resp. $\rho^{(k+1)} < \rho^{(k)}$).

Illustrative example.
Let $\mathcal{I} = \{1,2\}$, $\mathcal{A}_1^{(1)} = \mathcal{A}_1^{(2)} = \mathcal{A}_2^{(1)} = \mathcal{A}_2^{(2)} = \{1,2\}$ and the corresponding transition probabilities be given by the row vectors $p_i(f_i^{(1)}, f_i^{(2)}) = [p_{i1}(f_i^{(1)}, f_i^{(2)}), p_{i2}(f_i^{(1)}, f_i^{(2)})]$ for $f_i^{(1)}, f_i^{(2)} = 1,2$. The reward accrued in state i is equal to $r_i(f_i^{(1)}, f_i^{(2)})$.
The following example is borrowed from [5], Example 3.2.2, page 96. Let transition data and one-stage rewards be:

$p_1(1,1) = [0.5; 0.5]$	$r_1(1,1) = 10$	$p_1(1,2) = [0.5; 0.5]$	$r_1(1,2) = -6$
$p_1(2,1) = [0.8; 0.2]$	$r_1(2,1) = -4$	$p_1(2,2) = [0.8; 0.2]$	$r_2(2,2) = 8$
$p_2(1,1) = [0.3; 0.7]$	$r_2(1,1) = -2$	$p_1(1,2) = [0.3; 0.7]$	$r_2(1,2) = 5$
$p_2(2,1) = [0.9; 0.1]$	$r_2(2,1) = 4$	$p_2(2,2) = [0.9; 0.1]$	$r_2(2,2) = -10$

Considering the risk-sensitive model, we replace one-stage reward $r_i(f_i^{(1)}, f_i^{(2)})$ by
$\bar{r}_i(f_i^{(1)}, f_i^{(2)}) := \ln[r_i(f_i^{(1)}, f_i^{(2)})]$ if $r_i(f_i^{(1)}, f_i^{(2)}) > 0$ or by
$\bar{r}_i(f_i^{(1)}, f_i^{(2)}) := \ln[-r_i(f_i^{(1)}, f_i^{(2)})]$ if $r_i(f_i^{(1)}, f_i^{(2)}) < 0$.
Observe that $\mathrm{e}^{\gamma \bar{r}_i(f_i^{(1)}, f_i^{(2)})} = |r_i(f_i^{(1)}, f_i^{(2)})|^{\gamma}$. On recalling that
$q_{ij}(f_i^{(1)}, f_i^{(2)}) =: p_{ij}(f_i^{(1)}, f_i^{(2)}) \times r_i(f_i^{(1)}, f_i^{(2)})$, let the row vectors
$q_i(f_i^{(1)}, f_i^{(2)}) =: [q_{i1}(f_i^{(1)}, f_i^{(2)}), q_{i2}(f_i^{(1)}, f_i^{(2)})]$. Then $Q(f^{(1)}, f^{(2)})$ is the square (nonnegative) matrix whose i-th row is equal to $q_i(f_i^{(1)}, f_i^{(2)})$.

In particular, if $\gamma = 1$, resp. $\gamma = 0.5$,

$\gamma = 1$	$\gamma = 1$		$\gamma = 0.5$	$\gamma = 0.5$
$q_1(1,1) = [5; 5]$	$q_1(1,2) = [3; 3]$		$q_1(1,1) = [1.581; 1.581]$	$q_1(1,2) = [1.225; 1.225]$
$q_1(2,1) = [3.2; 0.8]$	$q_1(2,2) = [6.4; 1.6]$		$q_1(2,1) = [1.6; 0.4]$	$q_1(2,2) = [2.2628; 0.5656]$
$q_2(1,1) = [0.6; 1.4]$	$q_2(1,2) = [1.5; 3.5]$		$q_2(1,1) = [0.4242; 0.9899]$	$q_2(1,2) = [0.671; 1.5652]$
$q_2(2,1) = [3.6; 0.4]$	$q_2(2,2) = [9; 1]$		$q_2(2,1) = [1.8; 0.2]$	$q_2(2,2) = [2.846; 0.3162]$

As we can see, if $\gamma = 1$, on selecting in state 1 decision (1,1) and in state 2 decision (2,2) spectral radius of the resulting matrix is equal to 10 – maximum possible value. Similarly, selecting in state 1 decision (2,1) and in state 2 decision (1,1) spectral radius of the resulting matrix is equal to 3.4358 – minimum possible eigenvalue.
However, if $\gamma = 0.5$, on selecting in state 1 decision (1,1) and in state 2 decision (2,2) spectral radius of the resulting matrix is equal to 3.1621 – maximum possible value. Minimum possible value of the spectral radius is again obtained on selecting in state 1 decision (2,1) and in state 2 decision (1,1) spectral radius of the resulting matrix is equal to 1.8075, very close to spectral radius 1.8378 obtained if in state 1 decision (1,2) is selected and in state 2 decision (1,1) is unchanged.
Obviously, if $\gamma = 0$ the spectral radius equals one for all decisions.

Moreover, if the risk-sensitive coefficient $\gamma = 1$ by a direct calculation we can see that if the second players selects decision 2 (resp.1) in both states the first player maximize the profit by selecting action 2 in both states (resp. action 1 in state 1 and action 2 in state 2). If the second players selects decision 2 in state 1 and decision 1 in state 2 the optimal policy of the player 1 is to select action 2 in states 1 and 2. Finally, if the second players selects decision 1 in state 1 and decision 2 in state 2 the optimal policy of the player 1 is to select in state 1 action 1 and action 2 in state 2. Observe that in this case it is necessary to solve 4 problems concerning finding optimal policy of a risk-sensitive Markov decision chain.

To this end we suggest the following algorithmic procedure. More details and some numerical examples can be found in [13]. Observe that we restrict only on non-randomized decisions.

Algorithm 2. (Policy iterations for approximating optimal average reward.)

Step 0. Find matrix $Q^{(0)} := Q(f^{(1),0}, f^{(2),0})$ with $f^{(1),0} \in \mathcal{F}^{(1)}$, $f^{(2),0} \in \mathcal{F}^{(2)}$ such that its spectral radius is maximal (resp. minimal).

Step 1. For matrix $Q^{(k)}$ $(k = 0, 1, \ldots)$ calculate its spectral radius $\rho^{(k)}$ along with its right Perron eigenvector $v^{(k)}$.

Step 2. Construct (if possible) matrix $Q^{(k+1)} := Q(f^{(1),k+1}, f^{(2),k+1})$ with $f^{k+1} := (f^{(1),k+1}, f^{(2),k+1})$ where $f^{(1),k+1} \in \mathcal{F}^{(1)}$, $f^{(2),k+1} \in \mathcal{F}^{(2)}$, such that $f^{(1),k+1} = f^{(1),k}$ for k odd, resp. $f^{(2),k+1} = f^{(2),k}$ for k even, and

$$Q^{(k+1)} \cdot v^{(k)} \le \rho^{(k)}\, v^{(k)} = Q^{(k)} \cdot v^{(k)} \quad \text{if } k \text{ is odd} \qquad \text{resp.} \tag{20}$$

$$Q^{(k+1)} \cdot v^{(k)} \ge \rho^{(k)}\, v^{(k)} = Q^{(k)} \cdot v^{(k)} \quad \text{if } k \text{ is even} \tag{21}$$

Step 3. If for some $\ell = 0, 1, \ldots k$ it happens that $Q^{(k+1)} = Q^{(\ell)}$ then go to Step 4, else set $k := k + 1$ and repeat Step 1.

Step 4. Set $\bar{Q} := Q^{(\ell)} Q^{(\ell+1)} \cdot \ldots Q^{(k)}$. Calculate $\bar{\rho}$, the spectral radius of $\bar{Q}$ and stop.

Then $\rho^* = (\bar{\rho})^{\frac{1}{k-\ell}}$ is equal to the long-run risk-sensitive average reward generated by decisions of the first and second player. in the class of non-randomized policies.

Acknowledgements

This research was supported by the Czech Science Foundation under Grant 13-14445S and by CONACyT (Mexico) and AS CR (Czech Republic) under Project 171396.

References

[1] Bellman, R.: On a quasi-linear equation, *Canad. J. of Math.* **8** (1956), 198-202.

[2] Bellman, R.: A Markovian decision process, *J. of Math. and Mech.* **6** (1957), 679-684.

[3] Bellman, R.: *Dynamic Programming.* Princeton Univ. Press, Princeton. Princeton 1957.

[4] Blackwell, D.: Discrete dynamic programming, *Ann. Mathem. Statist.* **33** (1962), 719-726.

[5] Filar, J.A. and Vrieze, O.J.: *Competitive Markov Decision Processes.* Springer Verlag, Berlin, 1996.

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

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

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

[9] Kolokoltsov, V.N. and Malafayev, O.A.: *Understanding Game Theory.* World Scientific, Singapore, 2010.

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

[11] Nash, J.: Equilibrium points in n-person games, *Proc. Nat. Acad. Sci. U.S.A.* **36** (1950), 48–49.

[12] Shapley, L.S.: Stochastic games, *Proc. National Academy of Sciences U.S.A.* **39** (1953), 1095–1100.

[13] Sladký, K., Martínez-Cortés, V.M.: *Risk-Sensitive Optimality in Markov Games.* Research Report No.2361, Institute of Information Theory and Automation of the CAS, Prague 2017.