Math Meth Oper Res (2002) 56:473–479

Mathematical Methods
of Operations Research

Solution to the risk-sensitive average optimality equation in communicating Markov decision chains with finite state space: An alternative approach*

Rolando Cavazos-Cadena[1] **and Daniel Hernández-Hernández**[2]

[1] Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo Coah. 25315, MÉXICO (e-mail: rcavazos@narro.uaaan.mx)
[2] Centro de Investigación en Matemáticas, Apartado Postal 402, Guanajuato, Gto., 36000, MÉXICO (e-mail: dher@cimat.mx)

Manuscript received: November 2001/Final version received: April 2002

Abstract. This note concerns Markov decision chains with finite state and action sets. The decision maker is assumed to be risk-averse with constant risk sensitive coefficient λ, and the performance of a control policy is measured by the risk-sensitive average cost criterion. In their seminal paper Howard and Matheson established that, when the whole state space is a communicating class under the action of each stationary policy, then there exists a solution to the optimality equation for every $\lambda > 0$. This paper presents an alternative proof of this fundamental result, which explicitly highlights the essential role of the communication properties in the analysis of the risk-sensitive average cost criterion.

AMS Subject Classification: 93E20, 93B36

Key words: Contractive operator, Vanishing discount approach, Risk sensitive control

1 Introduction

This note concerns discrete-time Markov decision chains with finite state and action sets. The decision maker grades a random cost according to an exponential utility function with constant risk sensitivity λ, and it is supposed that she has an averse attitude with respect to risk; this feature is reflected in the positive sign of λ (see Section 2 for details). When the performance of a control strategy is measured by the risk sensitive (long-run) expected average cost criterion, Howard and Matheson proved in [6] that, if under the action of each stationary policy the state process is a completely communicating and aperi-

* This work was supported by the PSF Organization under Grant No. 010/300/01-4 and Conacyt Grant 37643-E

odic Markov chain, then the optimality equation has a solution for *arbitrary risk sensitivity coefficient* $\lambda > 0$. Their results rely on the Perron-Frobenious theory for maximum eigenvalues of positive matrices. On the other hand, it was recently shown that under the simultaneous Doeblin condition (ensuring that the Markov chain associated to each stationary policy has a single recurrent class), a solution to the risk-sensitive average optimality equation can be generally guaranteed *only when* λ *sufficiently small* (see [3], [5]), establishing a contrast with the original results by Howard and Matheson, and showing that the communication properties play a central role in the analysis of the risk-sensitive average cost criterion. *The objective of this note* is to provide an alternative proof of the original existence result by Howard and Matheson, which explicitly emphasizes the importance of having communication between every pair of states under the action of each stationary policy. The main idea consists in using a contractive operator, previously introduced in [4], and then parallel the so called "vanishing discount approach" in the study of the risk-neutral average cost criterion [1], [7], [8].

The organization of the paper is as follows: In Section 2 the decision model is briefly described, and the main result in [6] is stated in the form of Theorem 2.1. Next, a contractive operator is introduced in Section 3 and, finally, the proof of Theorem 2.1 is presented in Section 4.

Notation. Throughout the remainder $\mathbb{N}$ and $\mathbb{R}$ stand for the set of nonnegative integers and real numbers, respectively. If S is a finite set, $\mathscr{B}(S)$ denotes the class of real-valued functions defined on S, and for each $C \in \mathscr{B}(S)$, $\|C\| := \max_s |C(s)|$ denotes the corresponding maximum norm.

2 Decision model and the existence result

The Markov decision process (MDP) model M is specified by the four-tuple $M = \langle S, A, C, P \rangle$, where the state space S and the action sets A are *finite*, $C : S \times A \to \mathbb{R}$ is the cost function, and $P = [p_{x,y}(\cdot)]$ is the controlled transition law. This model is interpreted as follows: At each time $t \in \mathbb{N}$ the state $X_t = x \in S$ of a dynamical system is observed and an action $A_t = a \in A$ is applied. Then a cost $C(x, a)$ is incurred and, regardless of the previous states and actions, the state of the system at time $t+1$ will be $X_{t+1} = y \in S$ with probability $p_{x,y}(a)$; this is the Markov property of the decision model. Notice that it is assumed that every $a \in A$ is an admissible action at each state. As noted in [2], this condition does not imply any loss of generality.

The class $\mathscr{P}$ of admissible control policies consists of all the rules for choosing actions, which may depend on the current state and on the record of previous states and actions. Given the policy $\pi \in \mathscr{P}$ used to drive the system, and the initial state $X_0 = x$, the distribution of the state-action process $\{(X_t, A_t)\}$ is uniquely determined via Ionescu Tulcea's theorem (see, for instance, [1], [7], [8], for details). Such a distribution is denoted by P_x^π whereas E_x^π stands for the corresponding expectation operator. Define the set $\mathbb{F} := \prod_{x \in S} A$, which consists of all functions $f : S \to A$. A policy π is stationary if there exists $f \in \mathbb{F}$ such that, under π, at each time $t \in \mathbb{N}$ the action applied is $A_t = f(X_t)$. The class of stationary policies is naturally identified with $\mathbb{F}$, and with this convention $\mathbb{F} \subset \mathscr{P}$.

Performance Index and Utility Function. Let $\lambda > 0$ be a fixed number, here-

after referred to as the risk sensitivity coefficient, and define the corresponding utility function U_λ by

$$U_\lambda(x) = e^{\lambda x}, \quad x \in S. \tag{2.1}$$

At each transition step the decision maker grades a (bounded) random cost Y according to the expctation of $U_\lambda(Y)$. The certain equivalent $E(\lambda, Y)$ of Y is the unique real number satisfying $U_\lambda(E(\lambda, Y)) = E[U_\lambda(Y)]$, so that the controller is indifferent between incurring the random cost Y or paying the certain equivalent for sure; notice that

$$E(\lambda, Y) = \frac{1}{\lambda} \log(E[e^{\lambda Y}]).$$

Since $E(\lambda, Y) \geq E[Y]$, by Jensen's inequality, the controller is referred to as risk-averse. Suppose now that the system is driven by policy $\pi \in \mathscr{P}$ starting at $x \in S$, and let $J_n(\pi, x)$ be the certain equivalent of the total random cost incurred up to time n:

$$J_n(\pi, x) = \frac{1}{\lambda} \log(E_x^\pi[e^{\lambda \sum_{t=0}^n C(X_t, A_t)}]), \tag{2.2}$$

whereas the long-run average cost per statge is defined by

$$J(\pi, x) = \limsup_{n \to \infty} \frac{1}{n+1} J_n(\pi, x). \tag{2.3}$$

The (λ-)optimal average cost at state x is

$$J^*(x) = \inf_{\pi \in \mathscr{P}} J(\pi, x), \tag{2.4}$$

and a policy π is optimal if $J(\pi, x) = J^*(x)$ for each $x \in S$.

Optimality Equation. The optimality equation (OE) associated to the criterion in (2.2)–(2.4) is given by

$$U_\lambda(g + h(x)) = \min_a \left[\sum_y p_{xy}(a) U_\lambda(C(x, a) + h(y)) \right], \quad x \in S, \tag{2.5}$$

where g is a real number and $h : S \to \mathbb{R}$ is a given function. When the pair $(g, h(\cdot))$ satisfies this equality, it follows that: (i) The optimal average cost is g regardless of the initial state, i.e., $J^*(\cdot) \equiv g$, and (ii) If $f \in \mathbb{F}$ satisfies that, for each $x \in S$, $f(x)$ minimizes the term within brackets in (2.5), then f is optimal. Under the assumption that each stationary policy induces an aperiodic Markov chain, Howard and Matheson proved, via the Perron-Frobenious theory of positive matrices, the following fundamental result [6] (rewards, instead of costs, were used in that paper).

Theorem 2.1. *Suppose that each stationary policy induces a Markov chain for which the state space is a communicating class, i.e.,*

C: *For every $x, y \in S$ and $f \in \mathbb{F}$, there exists an integer $n = n(x, y, f)$ such that $P_x^f[X_n = y] > 0$.*

In this case, for each $\lambda > 0$, there exists a pair $(g, h(\cdot))$ satisfying the OE (2.5).

When condition (C) above fails, but the Markov chains associated to stationary policies have always a single recurrent class, the existence of a solution to the OE (2.5) can be ensured *only if $|\lambda|$ is sufficiently small* (see [3], [5]), establishing a contrast with the results in [6], and highlighting the role of Condition (C) in the existence of solutions to the OE *for arbitrary $\lambda > 0$*. The key analytical tool employed in [6] was the Perron-Frobenious theory for maximum eigenvalues of positive matrices. The objective of the paper is to provide an alternative proof of Theorem 2.1 which explicitly emphasizes the role of condition *C* to ensure that, *for every $\lambda > 0$*, the OE (2.5) has a solution. The following lemma will be useful.

Lemma 2.1. *Assume that Condition (C) in the statement of Theorem 2.1 holds. Let $f \in \mathbb{F}$ be fixed, and suppose that $\mathscr{A}$ is a nonempty subset of the state space satisfying the following property:*

$$x \in \mathscr{A} \Rightarrow y \in \mathscr{A} \quad \text{if } p_{xy}(f(x)) > 0. \tag{2.6}$$

In this case $\mathscr{A} = S$.

Proof. Let $y \in S$ be arbitrary and pick $z \in \mathscr{A}$. By condition (C) there exist states x_i, $i = 1, 2, \ldots, n$ such that (a) $x_0 = z$ and $x_n = y$, and (b) $p_{x_i x_{i+1}}(f(x_i)) > 0$ for $i = 0, 1, 2, \ldots, n-1$. In this case, $x_i \in \mathscr{A}$ implies that $x_{i+1} \in \mathscr{A}$, by (2.6). Therefore, $x_0 = z \in \mathscr{A}$ yields that $x_n = y \in \mathscr{A}$, so that $\mathscr{A} = S$, since the state y is arbitrary. □

3 A discounted operator

The argument used to establish Theorem 2.1 in the following section is based on the contractive operator introduced in (3.1) below which, in the present risk-sensitive context, allows to follow the so called 'vanishing discount approach' used to study the risk-neutral average cost criterion. For each $\alpha \in (0, 1)$, the operator $T_\alpha : \mathscr{B}(S) \to \mathscr{B}(S)$ is determined as follows: Given $W \in \mathscr{B}(S)$, $T_\alpha W$ is implicitly specified by

$$U_\lambda([T_\alpha W](x)) = \min_a \left[\sum_y p_{xy}(a) U_\lambda(C(x, a) + \alpha W(y)) \right], \quad x \in S. \tag{3.1}$$

Remark 3.1. (i) Suppose that the decision maker selects a single action at time $t = 0$, incurring a cost $C(X_0, A_0)$, and paying a terminal cost $W(X_1)$ at time $t = 1$. If this latter figure is discounted at rate α, $C(X_0, A_0) + \alpha W(X_1)$ represents the value at time zero of the total cost incurred by the controller, and $[T_\alpha W](x)$ is the minimum certain equivalent of those discounted random total costs.

(ii) Operator T_α was used in [4] to establish, for values of λ sufficiently close

to the origin, the existence of solutions to (2.5) for MDPs satisfying the simultaneous Doeblin condition over a denumerable state space.

The basic properties of T_α are stated in the following lemma.

Lemma 3.1. *For each* $\alpha \in (0,1)$ *assertions (i)–(iii) below hold.*

(i) T_α *is a contraction operator on* $\mathscr{B}(S)$ *with coefficient* α, *i.e. for each* $V, W \in \mathscr{B}(S)$,

$$\|TV - TW\| \leq \alpha \|V - W\|.$$

(ii) *There exists a unique function* $V_\alpha \in \mathscr{B}(S)$ *such that* $T_\alpha V_\alpha = V_\alpha$, *and*
(iii) $\|(1-\alpha)V_\alpha\| \leq \|C\|$.

Proof. Let $V, W \in \mathscr{B}(S)$. Noting that the inequality $C(x,a) + \alpha V(y) \leq C(x,a) + \alpha W(y) + \alpha\|V - W\|$ is always valid, it follows that $U_\lambda(C(x,a) + \alpha V(y)) \leq e^{\alpha\lambda\|V-W\|} U_\lambda(C(x,a) + \alpha W(y))$ (see (2.1)). Then, (3.1) yields that for every $x \in S$

$$\begin{aligned} U_\lambda([T_\alpha V](x)) &= \min_a \left[\sum_y p_{xy}(a) U_\lambda(C(x,a) + \alpha V(y)) \right] \\ &\leq e^{\alpha\lambda\|V-W\|} \min_a \left[\sum_y p_{xy}(a) U_\lambda(C(x,a) + \alpha W(y)) \right] \\ &= e^{\alpha\lambda\|V-W\|} U_\lambda([T_\alpha W](x)), \end{aligned}$$

so that $[T_\alpha V](x) \leq [T_\alpha W](x) + \alpha\|V - W\|$. Interchanging the roles of V and W this leads to $|[T_\alpha V](x) - [T_\alpha W](x)| \leq \alpha\|V - W\|$, and hence part (i) follows, since $x \in S$ is arbitrary. The existence of the unique fixed point V_α follows from part (i) and, to conclude, observe that for every $x, y \in S$ and $a \in A$, $-\|C\| - \alpha\|V_\alpha\| \leq C(x,a) + \alpha V_\alpha(x) \leq \|C\| + \alpha\|V_\alpha\|$, so that

$$U_\lambda(-\|C\| - \alpha\|V_\alpha\|) \leq \sum_y p_{xy}(a) U_\lambda(C(x,a) + \alpha V_\alpha(y)) \leq U_\lambda(\|C\| + \alpha\|V_\alpha\|).$$

After taking the minimum with respect to $a \in A$ in this relation, part (ii) and (3.1) together imply that $U_\lambda(-\|C\| - \alpha\|V_\alpha\|) \leq U_\lambda(V_\alpha(x)) \leq U_\lambda(\|C\| + \alpha\|V_\alpha\|)$, so that $|V_\alpha(x)| \leq \|C\| + \alpha\|V_\alpha\|$. Since $x \in S$ is arbitrary, it follows that $\|V_\alpha\| \leq \|C\| + \alpha\|V_\alpha\|$, which is equivalent to $(1-\alpha)\|V_\alpha\| \leq \|C\|$. □

4 Proof of the existence result

In this section a proof of Theorem 2.1 will be given. The argument uses the fixed points $\{V_\alpha\}_{\alpha \in (0,1)}$ in Lemma 3.1, and relies heavily on Lemma 2.1. To begin with, notice that (3.1) and the equaility $T_\alpha V_\alpha = V_\alpha$ together yield

$$U_\lambda(V_\alpha(x)) = \min_a \left[\sum_y p_{xy}(a) U_\lambda(C(x,a) + \alpha V_\alpha(y)) \right], \quad x \in S, \alpha \in (0,1) \tag{4.1}$$

and, since S is finite, for each $\alpha \in (0,1)$ there exists $z_\alpha \in S$ such that

$$V_\alpha(z_\alpha) = \min_{x \in S} V_\alpha(x). \tag{4.2}$$

Define

$$g_\alpha = (1-\alpha) V_\alpha(z_\alpha), \quad \text{and} \quad h_\alpha(x) = V_\alpha(x) - V_\alpha(z_\alpha), \quad x \in S. \tag{4.3}$$

With this notation, it is not difficult to see that, for each $\alpha \in (0,1)$, (4.1) is equivalent to

$$U_\lambda(g_\alpha + h_\alpha(x)) = \min_a \left[\sum_y p_{xy}(a) U_\lambda(C(x,a) + \alpha h_\alpha(y)) \right], \quad x \in S, \alpha \in (0,1), \tag{4.4}$$

and the finiteness of the action set implies that there exists a policy $f_\alpha \in \mathbb{F}$ such that

$$U_\lambda(g_\alpha + h_\alpha(x)) = \sum_y p_{xy}(f_\alpha(x)) U_\lambda(C(x,a) + \alpha h_\alpha(y)), \quad x \in S; \tag{4.5}$$

On the other hand, observe that Lemma 3.1(iii) and (4.3) together imply that $|g_\alpha| \le \|C\|$, whereas $h_\alpha(\cdot) \ge 0$, by (4.2). Therefore, given a sequence $\{\alpha_n\}$ in $(0,1)$ increasing to 1, the finiteness of S and $\mathbb{F}$ allow to pick a subsequence, still denoted by $\{\alpha_n\}$, such that

$$f_{\alpha_n} \equiv f \in \mathbb{F}, \quad \text{and} \quad z_{\alpha_n} \equiv z \in S, \quad n \in \mathbb{N}, \tag{4.6}$$

and the following limits exists:

$$\lim_{n \to \infty} g_{\alpha_n} = g, \quad \text{and} \quad \lim_{n \to \infty} h_{\alpha_n}(x) = h(x) \in [0, \infty], \quad x \in S. \tag{4.7}$$

Proof of Theorem 2.1. It will be shown that the function $h(\cdot)$ defined in (4.7) is finite, and that the pair $(g, h(\cdot))$ satisfies (2.5). First, define $\mathscr{A} := \{x \,|\, h(x) < \infty\}$, and notice that state z in (4.6) belongs to $\mathscr{A}$; indeed, by (4.3), (4.6) and (4.7), $h(z) = 0$. Suppose now that $x \in \mathscr{A}$. Replacing α by α_n in (4.5) and taking limit as n goes to ∞ in the resulting equality it follows, via (4.6) and (4.7), that

$$\infty > U_\lambda(g + h(x)) = \sum_y p_{xy}(f(x)) U_\lambda(C(x, f(x)) + h(y)), \quad x \in S, \tag{4.8}$$

and this yields that $h(y) < \infty$ when $p_{xy}(f(x)) > 0$. Therefore, $x \in \mathscr{A}$ implies that $y \in \mathscr{A}$ if $p_{xy}(f(x)) > 0$, so that, by Lemma 2.1, $\mathscr{A} = S$, i.e., $h(\cdot)$ is a finite

function. To conclude, note that, for each $x \in S$, $a \in A$ and $n \in \mathbb{N}$, (4.4) allows to write

$$U_\lambda(g_{\alpha_n} + h_{\alpha_n}(x)) \leq \sum_y p_{xy}(a) U_\lambda(C(x,a) + \alpha_n h_{\alpha_n}(y))$$

so that, letting n going to infinity, (4.7) yields

$$U_\lambda(g + h(x)) \leq \sum_y p_{xy}(a) U_\lambda(C(x,a) + h(y)).$$

Since the pair $(x,a) \in S \times A$ is arbitrary, this inequality and (4.8) show that the finite function $h(\cdot)$ and the constant g in (4.7) satisfy the optimality equation (2.5). □

References

[1] Araphostatis A, Borkar VK, Fernández-Gaucherand E, Gosh MK, Marcus SI (1993) Discrete time controlled Markov processes with average cost criterioa: a survey. SIAM Journal on Control and Optimization 31:282–334

[2] Borkar VK (1984) On minimum cost per unit of time control of Markov chains. SIAM Journal on Control and Optimization 21:965–984

[3] Cavazos-Cadena R, Fernández-Gaucherand E (1999) Controlled Markov chains with risk-sensitive criteria: average cost, optimality equations, and optimal solutions. Mathematical Methods of Operations Research 49:299–324

[4] Cavazos-Cadena R, Fernández-Gaucherand E (1999) Markov decision processes with risk-sensitive average cost criterion: The discounted stochastic games approach. Submitted for publication

[5] Hernández-Hernández D, Marcus SI (1996) Risk sensitive control of Markov processes in countable state space. Systems & Control Letters 29:147–155. Corrigendum (1998) 34:105–106

[6] Howard RA, Matheson JE (1972) Risk sensitive Markov decision processes. Management Science 18:356–369

[7] Hernández-Lerma O (1988) Adaptive Markov control processes. Springer-Verlag, New York

[8] Puterman ML (1994) Markov decision processes. Wiley, New York