APPLICATIONES MATHEMATICAE
27,2 (2000), pp. 167–185

**R. CAVAZOS-CADENA (Saltillo)**
**R. MONTES-DE-OCA (México)**

# OPTIMAL STATIONARY POLICIES IN RISK-SENSITIVE DYNAMIC PROGRAMS WITH FINITE STATE SPACE AND NONNEGATIVE REWARDS

*Abstract.* This work concerns controlled Markov chains with finite state space and nonnegative rewards; it is assumed that the controller has a constant risk-sensitivity, and that the performance of a control policy is measured by a risk-sensitive expected total-reward criterion. The existence of optimal stationary policies is studied within this context, and the main result establishes the optimality of a stationary policy achieving the supremum in the corresponding optimality equation, whenever the associated Markov chain has a unique positive recurrent class. Two explicit examples are provided to show that, if such an additional condition fails, an optimal stationary policy cannot be generally guaranteed. The results of this note, which consider both the risk-seeking and the risk-averse cases, answer an extended version of a question recently posed in Puterman (1994).

**1. Introduction.** This work concerns finite-state Markov decision processes (MDP's) endowed with a special type of expected total-reward criterion; such a performance index considers the controller's attitude toward risk when picking actions leading to random rewards. In this note it is assumed that the decison maker has a constant risk-sensitivity in the sense of Pratt (1964), so that the corresponding utility function, determined up to

1991 *Mathematics Subject Classification*: 93E20, 90C40.
*Key words and phrases*: Markov decision processes, risk-sensitive expected total-reward criterion, risk-sensitive optimality equation, unichain property.
The work of R. Cavazos-Cadena was generously supported by the Consejo Nacional de Ciencia y Tecnología under Grant No. E 120.3336, and by the PSF Organization under Grant No. 30-250–98–01.
The work of the R. Montes-de-Oca was partially supported by Consejo Nacional de Ciencia y Tecnologia (CONACyT) under grant No.400-200-5-25159-E.

an increasing affin transformation (Fishburn (1970)), is of the exponential type given in (2.1) below, and a control policy is graded via the expected utility of the total rewards obtained over an infinite horizon.

The incorporation of the controller's risk-sensitivity to the performace index of a control policy can be traced back, at least, to Howard and Matheson (1972), where finite MDP's endowed with a long run risk-sensitive average criterion were considered. Interest in this criterion in more general frameworks has recently sparkled; see for instance, Fleming and Hernández-Hernández (1997), Cavazos-Cadena and Fernández-Gaucherand (1999), and the references therein. In particular, in the latter reference the risk-sensitive expected total-reward criterion, formally introduced in Section 2 below, was used to obtain solutions to the average reward optimality equation. On the other hand, there is a vast amount of literature on the expected total-reward criterion in the risk-neutral case, and it is well known that important differences exist between negative and positive dynamic programs, corresponding to a nonpositive and a nonnegative reward function, respectively. For instance, in the negative case, every stationary policy achieving the optimum on the right-hand side of the optimality equation is optimal, but this assertion is no longer valid for positive dynamic programs; for details see, e.g., Chapter 7 in Puterman (1994). Recently, Ávila-Godoy (1998) studied finite MDP's endowed with the risk-sensitive expected total-reward criterion and, for nonpositive rewards, she proved that a stationary policy achieving the maximum in the corresponding optimality equation is necessarily optimal, extending a result by Strauch (1966) to the risk-sensitive context.

The main objective of the paper is to study the existence of optimal stationary policies with respect to the risk-sensitive expected total-reward criterion for positive dynamic programs, i.e., when the reward function is nonnegative. The results in this direction can be summarized as follows: (i) If a stationary policy $f$ is obtained maximizing the right-hand side of the optimality equation, then $f$ is optimal whenever it induces a Markov chain with a unique positive recurrent class (the unichain property); however, (ii) via two examples, it is shown that if this property fails, the existence of an optimal policy cannot be generally ensured. These results answer the risk-sensitive version of a question posed on page 324 of Puterman (1994).

The organization of the paper is as follows: In Section 2 the decision model and the idea of (exponential) utility function are introduced, and the optimality equation associated with the risk-sensitive expected total reward is established in Section 3. Next, in Section 4 the main result of the paper is stated as Theorem 4.1, and the role played by the unichain property in the existence of optimal stationary policies is discussed. The proof of Theorem 4.1 is given in Section 6 after the preliminary results presented in Section 5. Finally, the paper concludes in Section 7 with some brief comments.

**2. Decision model.** Throughout the remainder $M = (S, A, \{A(x)\}, R, P)$ stands for the usual MDP, where the state space $S$ is *finite*, the metric space $A$ is the control (or action) set, and for each $x \in S$, $A(x) \subset A$ is the (nonempty and) measurable subset of admissible actions at state $x$; define the class of *admissible pairs* by $\mathbb{K} := \{(x, a) \mid a \in A(x),\ x \in S\}$. On the other hand, $R : \mathbb{K} \to \mathbb{R}$ is the reward function and $P = [p_{xy}(\cdot)]$ is the controlled transition law. This model $M$ has the following interpretation: At each time $t \in \mathbb{N} := \{0, 1, 2, \ldots\}$ the state of a dynamical system is observed, say $X_t = x \in S$, and an action $A_t = a \in A(x)$ is chosen. As a consequence, a reward $R(x, a)$ is earned and, regardless of which states and actions were observed and applied before $t$, the state of the system at time $t + 1$ will be $X_{t+1} = y \in S$ with probability $p_{xy}(a)$; this is the Markov property of the decision model.

ASSUMPTION 2.1. For every $x, y \in S$, the mappings $a \mapsto R(x, a)$ and $a \mapsto p_{xy}(a)$ are measurable on $A(x)$.

*Utility function.* Given $\lambda \in \mathbb{R}$, hereafter referred to as the (constant) *risk-sensitivity coefficient*, the corresponding utility function $U_\lambda : \mathbb{R} \to \mathbb{R}$ is defined as follows: For $x \in \mathbb{R}$,

$$U_\lambda(x) := \begin{cases} \operatorname{sign}(\lambda) e^{\lambda x} & \text{if } \lambda \neq 0, \\ x & \text{if } \lambda = 0; \end{cases} \tag{2.1}$$

notice that $U_\lambda(\cdot)$ is always a strictly increasing function, and

$$U_\lambda(c + x) = e^{\lambda c} U_\lambda(x), \quad \lambda \neq 0,\ x, c \in \mathbb{R}. \tag{2.2}$$

A controller with risk-sensitivity $\lambda$ grades a random reward $Y$ via the expectation of $U_\lambda(Y)$, so that if two decision strategies $\delta_1$ and $\delta_2$ lead to obtain random rewards $Y_1$ and $Y_2$, respectively, $\delta_1$ will be preferred if $E[U_\lambda(Y_1)] > E[U_\lambda(Y_2)]$, whereas the controller will be indifferent between $\delta_1$ and $\delta_2$ when $E[U_\lambda(Y_1)] = E[U_\lambda(Y_2)]$. Let $Y$ be a given random variable, and suppose that the expected value of $U_\lambda(Y)$ is well defined, a condition that is always valid when $\lambda \neq 0$. In this case, the *certain equivalent* of $Y$ with respect to $U_\lambda(\cdot)$ is defined by

$$E(\lambda, Y) = \begin{cases} \dfrac{1}{\lambda} \log(E[e^{\lambda Y}]), & \lambda \neq 0, \\ E[Y], & \lambda = 0, \end{cases} \tag{2.3}$$

where the usual conventions $\log(\infty) = \infty$ and $\log(0) = -\infty$ are in force; combining (2.1) and (2.3) it follows that

$$U_\lambda(E(\lambda, Y)) = E[U_\lambda(Y)]. \tag{2.4}$$

Thus, for a controller with risk-sensitivity $\lambda$, the opportunity of getting a random reward $Y$ can be fairly interchanged with obtaining the corre-

sponding certain equivalent $E(\lambda, Y)$ for sure. Suppose now that the random variable $Y$ is not constant. When $\lambda > 0$ the utility function $U_\lambda(\cdot)$ in (2.1) is convex, so that Jensen's inequality yields that $E(\lambda, Y) > E[Y]$; similarly, $E(\lambda, Y) < E[Y]$ if $\lambda < 0$, since in this case $U_\lambda(\cdot)$ is strictly concave. A decision maker grading a random reward $Y$ according to the certain equivalent $E(\lambda, Y)$ is referred to as *risk-seeking* if $\lambda > 0$, and *risk-averse* when $\lambda < 0$; if $\lambda = 0$, the decision maker is *risk-neutral*.

REMARK 2.1. (i) Notice that if $P[Y = c] = 1$ for some $c \in \mathbb{R}$, then $E(\lambda, Y) = c$.

(ii) Let $Y$ and $W$ be two random variables satisfying $P[Y \geq W] = 1$. Since $U_\lambda(\cdot)$ is increasing, it follows that $U_\lambda(E(\lambda, Y)) = E[U_\lambda(Y)] \geq E[U_\lambda(W)] = U_\lambda(E(\lambda, W))$, and then $E(\lambda, Y) \geq E(\lambda, W)$. In particular,

(iii) If $P[Y \geq 0] = 1$ (resp. $P[Y \leq 0] = 1$) then $E(\lambda, Y) \geq 0$ (resp. $E(\lambda, Y) \leq 0$).

*Policies.* For each $t \in \mathbb{N}$, the space of histories up to time $t$ is recursively defined by $\mathbb{H}_0 = S$, and $\mathbb{H}_t = \mathbb{K} \times \mathbb{H}_{t-1}$ for $t \geq 1$; a generic element of $\mathbb{H}_t$ is denoted by $h_t = (x_0, a_0, x_1, \ldots, x_{t-1}, a_{t-1}, x_t)$. A *policy* is a sequence $\{\pi_t \mid t \in \mathbb{N}\}$ such that $\pi_t$ is a stochastic kernel on $A$ given $\mathbb{H}_t$ of a special type; more precisely, for each $h_t \in \mathbb{H}_t$, $\pi_t(\cdot \,|\, h_t)$ is a probability measure on $\mathcal{B}(A)$, the space of Borel subsets of $A$, and it satisfies $\pi_t(A(x_t) \,|\, h_t) = 1$, whereas for each $B \in \mathcal{B}(A)$, the mapping $h_t \mapsto \pi_t(B \,|\, h_t)$ is a measurable function on $\mathbb{H}_t$. When the policy $\pi$ is used to pick the action to be applied at every decision time, $\pi_t(B \,|\, h_t)$ is the probability of the event $[A_t \in B]$ given the history of the decison process up to time $t$; throughout the remainder $\mathcal{P}$ denotes the class of all policies. Given the initial state $X_0 = x$ and the policy $\pi \in \mathcal{P}$ being used to drive the system, under Assumption 2.1 the distribution of the state-action process $\{(X_t, A_t)\}$ is uniquely determined via the Ionescu Tulcea's theorem (see, for instance, Hernández-Lerma (1989), Hinderer (1970) or Puterman (1994)); such a distribution is denoted by $P_\pi[\cdot \,|\, X_0 = x]$ whereas $E_\pi[\cdot \,|\, X_0 = x]$ stands for the corresponding expectation operator. Define $\mathbb{F} := \prod_{x \in S} A(x)$, so that $\mathbb{F}$ consists of all (choice) functions $f : S \to A$ satisfying $f(x) \in A(x)$ for all $x \in S$. A policy $\pi$ is *stationary* if there exists $f \in \mathbb{F}$ such that when $X_t = x$ is observed, the action prescribed by $\pi$ is *always* $f(x)$, i.e., $\pi_t(\{f(x_t)\} \,|\, h_t) = 1$ for every $h_t \in \mathbb{H}_t$ and $t \in \mathbb{N}$; the class of stationary policies is naturally identified with $\mathbb{F}$, and with this convention $\mathbb{F} \subset \mathcal{P}$. Notice, finally, that under the action of each policy $f \in \mathbb{F}$, the state process $\{X_t\}$ is a Markov chain with stationary transition mechanism [Ross (1970)].

*Performance index.* Throughout the remainder $\lambda$ stands for a nonzero real number, and the *$\lambda$-sensitive expected total reward* at state $x \in S$ under

policy $\pi \in \mathcal{P}$ is defined by

$$V_\lambda(\pi, x) = \frac{1}{\lambda} \log(E_\pi[e^{\lambda \sum_{t=0}^{\infty} R(X_t, A_t)} \mid X_0 = x]), \tag{2.5}$$

so that

$$U_\lambda(V_\lambda(\pi, x)) = E_\pi\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x\Big] \tag{2.6}$$

(see (2.3) and (2.4)). Thus, when the system is driven by the policy $\pi$ starting at $x$, $V_\lambda(\pi, x)$ is the certain equivalent of the total reward $\sum_{t=0}^{\infty} R(X_t, A_t)$ with respect to $U_\lambda(\cdot)$; the *$\lambda$-optimal value function* is

$$V_\lambda^*(x) = \sup_\pi V_\lambda(\pi, x), \tag{2.7}$$

and a policy $\pi$ is *$\lambda$-optimal* if $V_\lambda(\pi, x) = V_\lambda^*(x)$ for all $x \in S$. Although the expected value in (2.5) is always well defined, under Assumption 2.1 alone it can happen that $V_\lambda^*(x)$ is not finite for some $x \in S$; such an inconvenience is now excluded from the discussion.

ASSUMPTION 2.2. For each $x \in S$, $V_\lambda^*(x)$ is finite.

**3. The risk-sensitive optimality equation.** As already mentioned, the main objective of the paper is to study the existence of $\lambda$-optimal stationary policies, and the first step in this direction is to establish the optimality equation satisfied by the optimal value function $V_\lambda^*$.

LEMMA 3.1. *Under Assumptions* 2.1 *and* 2.2 *the optimal value function* $V_\lambda^*(\cdot)$ *in* (2.7) *satisfies the following* $\lambda$*-optimality equation* ($\lambda$*-OE*):

$$U_\lambda(V_\lambda^*(x)) = \sup_{a \in A(x)} \Big[e^{\lambda R(x,a)} \sum_y p_{xy}(a) U_\lambda(V_\lambda^*(y))\Big], \quad x \in S. \tag{3.1}$$

This result has been established in the literature under several conditions. In Cavazos-Cadena and Fernández-Gaucherand (1999), equation (3.1) was proved for models with denumerable state space under a simultaneous Doeblin condition, and then it was used to show the existence of solutions to the $\lambda$-sensitive average reward optimality equation. Also, Ávila-Godoy (1998) obtained the above $\lambda$-OE for models with finite state and action spaces. Since Lemma 3.1 plays a central role in the subsequent development, a detailed proof will be provided.

*Proof of Lemma 3.1.* Observe that (2.2) yields

$$U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) = e^{\lambda R(X_0, A_0)} U_\lambda\Big(\sum_{t=1}^{\infty} R(X_t, A_t)\Big),$$

so that, for arbitrary $\pi \in \mathcal{P}$, $x, x_1 \in S$ and $a_0 \in A(x)$, the Markov property implies

$$
\begin{aligned}
(3.2)\qquad & E_\pi\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x,\ A_0 = a_0,\ X_1 = x_1\Big] \\
&= e^{\lambda R(x,a_0)} E_\pi\Big[U_\lambda\Big(\sum_{t=1}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x,\ A_0 = a_0,\ X_1 = x_1\Big] \\
&= e^{\lambda R(x,a_0)} E_{\pi'}\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x_1\Big]
\end{aligned}
$$

where the shifted policy $\pi'$ is defined by $\pi'_t(\cdot \,|\, h_t) = \pi_{t+1}(\cdot \,|\, x, a_0, h_t)$. Combining the last equality with (2.6) and (2.7) gives

$$
\begin{aligned}
&E_\pi\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x,\ A_0 = a_0,\ X_1 = x_1\Big] \\
&\qquad = e^{\lambda R(x,a_0)} U_\lambda(V_\lambda(\pi', x_1)) \le e^{\lambda R(x,a_0)} U_\lambda(V^*_\lambda(x_1)),
\end{aligned}
$$

since $U_\lambda(\cdot)$ is increasing. Taking expectation with respect to $X_1$ yields

$$
\begin{aligned}
&E_\pi\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x, A_0 = a_0\Big] \\
&\qquad\qquad \le e^{\lambda R(x,a_0)} \sum_y p_{xy}(a_0) U_\lambda(V^*_\lambda(y)) \\
&\qquad\qquad \le \sup_{a \in A(x)} \Big[e^{\lambda R(x,a)} \sum_y p_{xy}(a) U_\lambda(V^*_\lambda(y))\Big],
\end{aligned}
$$

and then, taking the expected value with respect to $A_0$ and using (2.6), we get

$$
\begin{aligned}
U_\lambda(V_\lambda(\pi, x)) &= E_\pi\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x\Big] \\
&\le \sup_{a \in A(x)} \Big[e^{\lambda R(x,a)} \sum_y p_{xy}(a) U_\lambda(V^*_\lambda(y))\Big].
\end{aligned}
$$

Since the policy $\pi$ is arbitrary in this argument and $U_\lambda(\cdot)$ is increasing and continuous, the last inequality and (2.7) together yield

$$
(3.3)\qquad U_\lambda\Big(V^*_\lambda(x)\Big) \le \sup_{a \in A(x)} \Big[e^{\lambda R(x,a)} \sum_y p_{xy}(a) U_\lambda(V^*_\lambda(y))\Big].
$$

To establish the reverse inequality, fix $\varepsilon > 0$, and for each $x \in S$ select an arbitrary action $a_x \in A(x)$ and a policy $\pi^x i \in \mathcal{P}$ satisfying

$$
(3.4)\qquad V_\lambda(\pi^x, x) \ge V^*_\lambda(x) - \varepsilon;
$$

see (2.7).i Next, define a new policy $\pi$ as follows: For each state $x$, $\pi_0(\{a_x\}|x) = 1$, whereas for $h_t \in \mathbb{H}_t$ with $t \geq 1$,

$$\pi_t(\cdot \,|\, h_t) = \pi_{t-1}^{x_1}(\cdot \,|\, x_t, a_{t-1}, \ldots, x_2, a_1, x_1).$$

Thus, under $\pi$, the action $a_x$ is applied at time $t = 0$ whenever $X_0 = x$, and from time 1 onwards, if $X_1 = y$ actions are chosen according to $\pi^y$ as if the process started again. Notice that for this policy $\pi$, the shifted policy $\pi'$ in (3.2) is $\pi^{x_1}$ when $X_1 = x_1$, so that

$$\begin{aligned}
E_\pi\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x,\ A_0 = a_x,\ X_1 = x_1\Big] \\
&= e^{\lambda R(x,a_x)} E_{\pi^{x_1}}\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x_1\Big] \\
&= e^{\lambda R(x,a_x)} U_\lambda(V_\lambda(\pi^{x_1}, x_1)) \\
&\geq e^{\lambda R(x,a_x)} U_\lambda(V_\lambda^*(x_1) - \varepsilon) \\
&= e^{-\lambda\varepsilon} e^{\lambda R(x,a_x)} U_\lambda(V_\lambda^*(x_1)),
\end{aligned}$$

where the inequality used (3.4) and the fact that $U_\lambda(\cdot)$ is strictly increasing, and the last equality stems from (2.2). Taking expectation with respect to $A_0$ and $X_1$ and using (2.6) yields

$$\begin{aligned}
U_\lambda(V_\lambda(\pi, x)) &= E_\pi\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x\Big] \\
&\geq e^{-\lambda\varepsilon} e^{\lambda R(x,a_x)} \sum_y p_{xy}(a_x) U_\lambda(V_\lambda^*(y)).
\end{aligned}$$

Since $U_\lambda(V_\lambda^*(x)) \geq U_\lambda(V_\lambda(\pi, x))$, this inequality implies

$$U_\lambda(V_\lambda^*(x)) \geq e^{-\lambda\varepsilon} e^{\lambda R(x,a_x)} \sum_y p_{xy}(a_x) U_\lambda(V_\lambda^*(y)),$$

and then, since $\varepsilon > 0$ and $a_x \in A(x)$ are arbitrary, it follows that

$$U_\lambda(V_\lambda^*(x)) \geq \sup_{a \in A(x)} \Big[e^{\lambda R(x,a_x)} \sum_y p_{xy}(a_x) U_\lambda(V_\lambda^*(y))\Big],$$

and the desired conclusion is obtained by combining this inequality and (3.3). ∎

The $\lambda$-OE in (3.1) is an important tool to study the existence of $\lambda$-optimal stationary policies. Recently, Ávila-Godoy (1998) proved that, when the reward function satisfies $R(\cdot, \cdot) \leq 0$, if a stationary policy $f$ is such that $f(x)$ achieves the optimum on the right-hand side of (3.1) for each $x \in S$,

then $f$ is $\lambda$-optimal, providing an extension of a result established by Strauch (1966) in risk-neutral dynamic programming; see also Puterman (1994).

THEOREM 3.1 [Ávila-Godoy (1998)]. *Suppose that Assumptions* 2.1 *and* 2.2 *are valid and that* $R(\cdot,\cdot) \leq 0$. *Let* $f \in \mathbb{F}$ *be such that, for every* $x \in S$, $f(x)$ *is a maximizer of the term in brackets on the right-hand side of the* $\lambda$*-OE. In this case,* $f$ *is* $\lambda$*-optimal.*

REMARK 3.1. Suppose that the word *measurable* in Assumption 2.1 is replaced by *continuous*, and that this modified Assumption 2.1 as well as Assumption 2.2 hold. Since the state space is finite, it follows that for each $x \in S$, $a \mapsto e^{\lambda R(x,a)} \sum_y p_{xy}(a) U_\lambda(V_\lambda^*(y))$ is a continuous mapping. Hence, when the action sets are compact, it follows that there exists a policy $f \in \mathbb{F}$ such that, for every $x \in S$, $f(x)$ maximizes this mapping; by Theorem 3.1, such a policy is $\lambda$-optimal whenever the reward function is nonpositive.

**4. Risk-sensitive positive dynamic programs.** This section concerns the existence of $\lambda$-optimal stationary policies in dynamic programs with nonnegative rewards. The main result of this note, stated below as Theorem 4.1, provides an additional sufficient condition to ensure the $\lambda$-optimality of a stationary policy $f$ achieving the optimum on the right-hand side of the $\lambda$-OE.

THEOREM 4.1. *Suppose that Assumptions* 2.1 *and* 2.2 *hold and that* $R(\cdot,\cdot) \geq 0$. *Assume that an* $f \in \mathbb{F}$ *satisfies*:

(i) *For each* $x \in S$,

$$U_\lambda(V_\lambda(x)) = e^{\lambda R(x,f(x))} \sum_y p_{xy}(f(x)) U_\lambda(V_\lambda^*(y)), \quad x \in S. \tag{4.1}$$

(ii) $f$ *has the unichain property, i.e., the Markov chain induced by* $f$ *has a unique positive recurrent class.*

*In this case,* $f$ *is* $\lambda$*-optimal.*

This result will be proved in Section 6, after presenting the necessary technical tools in the following section; at this moment, observe that the existence of a policy $f$ satisfying (4.1) can be guaranteed under the continuity-compactness conditions in Remark 3.1 but, in contrast to the conclusion of Theorem 3.1, for nonnegative rewards the $\lambda$-optimality of such a policy is asserted only if the corresponding Markov chain has a unique positive recurrent class. The remainder of this section concerns the role of the unichain property in the existence of $\lambda$-optimal stationary policies but, before going directly over this point, it is convenient to establish the following characterization of the optimal value function $V_\lambda^*(\cdot)$, which will be very useful

in the analysis of Examples 4.1 and 4.2 below, as well as in the proof of Theorem 4.1.

LEMMA 4.1. *Suppose that Assumption* 2.1 *is valid and that* $R \geq 0$. *Let* $W : S \to [0, \infty)$ *be a function satisfying*

$$U_\lambda(W) \geq \sup_{a \in A(x)} \Big[ e^{\lambda R(x,a)} \sum_y p_{xy}(a) U_\lambda(W(y)) \Big], \quad x \in S. \tag{4.2}$$

*In this case* $W \geq V_\lambda^*$.

P r o o f. Notice that (4.2) implies that the inequality

$$U_\lambda(W(x)) \geq E_\pi[e^{\lambda R(X_0, A_0)} U_\lambda(W(X_1)) \,|\, X_0 = x]$$

is always valid, and then, using the Markov property, an induction argument yields that for every $\pi \in \mathcal{P}$, $x \in S$ and $n \in \mathbb{N}$,

$$U_\lambda(W(x)) \geq E_\pi[e^{\lambda \sum_{t=0}^n R(X_t, A_t)} U_\lambda(W(X_{n+1})) \,|\, X_0 = x].$$

Since $W(\cdot) \geq 0$ and $U_\lambda(\cdot)$ is increasing, it follows that $U_\lambda(W(X_{n+1})) \geq U_\lambda(0)$, so that

$$\begin{aligned} U_\lambda(W(x)) &\geq E_\pi[e^{\lambda \sum_{t=0}^n R(X_t, A_t)} U_\lambda(0) \,|\, X_0 = x] \\ &= E_\pi \Big[ U_\lambda\Big( \sum_{t=0}^n R(X_t, A_t) \Big) \,\Big|\, X_0 = x \Big], \end{aligned} \tag{4.3}$$

where (2.2) was used to obtain the equality. Consider now the following two cases:

CASE 1: $\lambda > 0$. In this situation

$$0 \leq U_\lambda\Big( \sum_{t=0}^n R(X_t, A_t) \Big) \nearrow U_\lambda\Big( \sum_{t=0}^\infty R(X_t, A_t) \Big),$$

since $R(\cdot, \cdot) \geq 0$. Thus, the monotone convergence theorem implies, after taking the limit as $n$ goes to $\infty$ on the right-hand side of (4.3), that for every $x \in S$,

$$U_\lambda(W(x)) \geq E_\pi \Big[ U_\lambda\Big( \sum_{t=0}^\infty R(X_t, A_t) \Big) \,\Big|\, X_0 = x \Big] \tag{4.4}$$

or, equivalently,

$$U_\lambda(W(x)) \geq U_\lambda(V_\lambda(\pi, x)), \quad x \in S,\ \pi \in \mathcal{P}; \tag{4.5}$$

see (2.6).

Case 2: $\lambda < 0$. Under this condition, the nonnegativity of the reward function and the definition of $U_\lambda(\cdot)$ in (2.1) together yield that

$$U_\lambda(R(X_0, A_0)) \le U_\lambda\Big(\sum_{t=0}^{n} R(X_t, A_t)\Big) \nearrow U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \le 0.$$

Then (4.3) implies, via the dominated convergence theorem, that (4.4) and (4.5) are also valid for $\lambda < 0$.

To summarize, inequality (4.5) has been established for every $\lambda \neq 0$; consequently, since $U_\lambda(\cdot)$ is increasing, $W(x) \ge V_\lambda(\pi, x)$ for every $x \in S$ and $\pi \in \mathcal{P}$, and then (2.7) yields that $W(\cdot) \ge V_\lambda^*(\cdot)$. ∎

The following examples, built upon ideas presented in Strauch (1966) and Cavazos-Cadena and Montes-de-Oca (1999), use Lemma 4.1 to show explicitly that for nonnegative rewards the $\lambda$-optimality of a policy satisfying (4.1) cannot be generally ensured if the unichain property fails. First, the risk-seeking case $\lambda > 0$ is analyzed.

Example 4.1. Let $\lambda > 0$ be a fixed risk-sensitivity coefficient, and consider an MDP with state and action spaces given by $S = \{0, 1\}$ and $A = [0, 1]$, respectively, whereas the action sets are $A(1) = [0, 1]$ and $A(0) = \{0\}$. Define the transition law and the reward function by

$$p_{00}(0) = 1, \quad p_{11}(a) = a = 1 - p_{10}(a), \quad a \in [0, 1]; \tag{4.6}$$

$$R(0, 0) = 0, \quad R(1, a) \equiv r(a) = \frac{1 - a}{2\lambda}, \quad a \in [0, 1]. \tag{4.7}$$

From the specifications in this example, it is clear that under the action of every policy $\pi \in \mathcal{P}$, state 0 is absorbing and $P_\pi[\sum_{t=0}^{\infty} R(X_t, A_t) = 0 \,|\, X_0 = 0] = 1$. Therefore, $E_\pi[U_\lambda(\sum_{t=0}^{\infty} R(X_t, A_t)) \,|\, X_0 = 0] = 1$, so that $V_\lambda(\pi, 0) = 0$ (by Remark 2.1). Hence, $V_\lambda^*(0) = 0$; see (2.5)–(2.7). On the other hand, the stationary policies are naturally indexed by the action prescribed at state 1: $f_a \in \mathbb{F}$ is given by

$$f_a(1) = a, \quad f_a(0) = 0, \quad a \in [0, 1]. \tag{4.8}$$

In the following proposition, the $\lambda$-optimal expected total reward is determined, and it is shown that no optimal stationary policy exists.

Proposition 4.1. *For Example* 4.1, *the following assertions are valid*:

(i) *The expected total reward at state* 1 *under policy* $f_a$ *is determined by*

$$\lambda V_\lambda(1, f_a) = \begin{cases} \log\left(\dfrac{1 - a}{e^{-(1-a)/2} - a}\right) & \text{if } a \in [0, 1), \\ 0 & \text{if } a = 1. \end{cases}$$

(ii) *The mapping* $a \mapsto V_\lambda(f_a, 1)$ *is increasing in* $a \in [0, 1)$.

(iii) *As* $a \nearrow 1$, $V_\lambda(f_a, 1) \nearrow L$, *where* $L = \lambda^{-1} \log(2)$.

*Moreover*,

(iv) $L = V_\lambda^*(1)$.

*Consequently*,

(v) *A $\lambda$-optimal stationary policy does not exist.*

P r o o f. (i) First, notice that state 1 is absorbing under policy $f_1$. Since $R(1,1) = 0$ (see (4.7)), it follows that if $X_0 = 1$ and the system is driven by policy $f_1$, then a reward zero is earned forever, so that $V_\lambda(f_1, 1) = 0$. To complete the proof of part (i), fix $a \in [0, 1)$ and notice

$$ae^{\lambda r(a)} = ae^{(1-a)/2} < 1. \tag{4.9}$$

In fact, the strict concavity of the logarithmic function yields $\log(a) < a - 1 < (a-1)/2$ (since $a \in [0,1)$), so that $\log(a) + (1-a)/2 < 0$, which is equivalent to (4.9). Next, let $T = \min\{n > 0 \mid X_n = 0\}$ be the first return time to state 0. From the specification of the transition law in (4.6), it is clear that when the system is driven by $f_a$ and the initial state is $X_0 = 1$, $T$ has a geometric distribution given by

$$P_{f_a}[T = k \,|\, X_0 = 1] = a^{k-1}(1-a), \qquad k = 1, 2, \ldots,$$

whereas a reward $r(a)$ will be earned while the system stays at state 1. Hence,

$$\begin{aligned} U_\lambda(V_\lambda(f_a, 1)) &= E_{f_a}\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = 1\Big] \\ &= E_{f_a}\Big[U_\lambda\Big(\sum_{t=0}^{T-1} R(X_t, A_t)\Big) \,\Big|\, X_0 = 1\Big] \\ &= E_{f_a}[U_\lambda(Tr(A_t)) \,|\, X_0 = 1] = E_{f_a}[U_\lambda(0)e^{\lambda T r(A_t)} \,|\, X_0 = 1] \\ &= U_\lambda(0) \sum_{k=1}^{\infty} e^{\lambda k r(a)} a^{k-1}(1-a) = U_\lambda(0) \frac{e^{\lambda r(a)}(1-a)}{1 - ae^{\lambda r(a)}}; \end{aligned}$$

notice that the geometric series in this argument is convergent, by (4.9). Using the definition of $U_\lambda(\cdot)$ and $r(\cdot)$ in (2.1) and (4.7), respectively, shows that

$$\lambda V_\lambda(f_a, 1)) = \log\left(\frac{e^{\lambda r(a)}(1-a)}{1 - ae^{\lambda r(a)}}\right) = \log\left(\frac{(1-a)}{e^{-(1-a)/2} - a}\right).$$

(ii) From part (i), it follows that for $a \in [0, 1)$,

$$\lambda \frac{dV_\lambda(f_a, 1)}{da} = e^{-(1-a)/2} \frac{e^{(1-a)/2} - 1 - (1-a)/2}{(1-a)(e^{-(1-a)/2} - a)}, \tag{4.10}$$

where the derivative is from the right at $a = 0$. Since $a \in [0, 1)$, (4.9) yields that the denominator in this expression is positive. Also, the strict

convexity of the exponential function yields that $e^x - 1 - x > 0$ for every $x \neq 0$. Therefore the numerator in (4.10) is also positive for every $a \in [0,1)$, and then $\lambda dV_\lambda(f_a,1)/da > 0$; this yields the conclusion, since $\lambda$ is positive.

(iii) The assertion follows by combining part (i) and the L'Hospital rule.

(iv) Define $W(1) = L$ and $W(0) = 0$. In this case, for arbitrary $a \in [0,1)$, parts (ii) and (iii) yield $W(1) > V_\lambda(f_a,1)$ and since $U_\lambda(\cdot)$ is strictly increasing, it follows that

$$U_\lambda(W(1)) > U_\lambda(V_\lambda(f_a,1)) = \frac{(1-a)e^{(1-a)/2}}{1-ae^{(1-a)/2}}.$$

After some rearrangements using (4.6) and (4.7), this yields that for every $a \in [0,1)$,

$$U_\lambda(W(1)) > e^{\lambda R(1,a)}[p_{11}(a)U_\lambda(W(1)) + p_{10}(0)U_\lambda(W(0))].$$

Since $p_{11}(1) = 1$ and $R(1,1) = 0$, this relation turns into equality for $a = 1$, so that

$$U_\lambda(W(1)) = \sup_{a\in A(1)} [e^{\lambda R(1,a)}\{p_{11}(a)U_\lambda(W(1)) + p_{10}(0)U_\lambda(W(0))\}].$$

On the other hand, as $p_{00}(0) = 1$ and $R(0,0) = 0$, it follows that

$$e^{\lambda R(0,0)}[p_{00}(0)U_\lambda(W(0)) + p_{01}(0)U_\lambda(W(1))] = U_\lambda(W(0)),$$

and then the nonnegative function $W(\cdot)$ satisfies the $\lambda$-OE. Therefore, $L = W(1) \geq V_\lambda^*(1)$ by Lemma 4.1, and then part (iii) implies that $L = V_\lambda^*(1)$.

(v) This part follows from (i)–(iv). ∎

In Example 4.1 Assumptions 2.1 and 4.2 are valid; moreover, the continuity-compactness conditions of Remark 3.1 hold, so that there exists a policy $f \in \mathbb{F}$ satisfying (4.1). However, such a policy is not $\lambda$-optimal, by Proposition 4.1. According to Theorem 4.1, $f$ does not have the unichain property. In fact, it is not difficult to verify directly that the unique policy satisfying (4.1) is $f_1$, under which the sets $\{0\}$ and $\{1\}$ are closed, i.e., $f_1$ is not unichain. The following example considers a risk-averse controller.

Example 4.2. Given a fixed negative risk-sensitivity coefficient $\lambda$, consider an MDP whose state and action spaces, as well as the sets of admissible actions are the same as in Example 4.1. Let the transition law be given as in (4.6), and define the reward function as follows:

$$R(0,0) = 0, \qquad R(1,a) \equiv r(a) = \frac{1-a}{(1-a/2)|\lambda|}, \qquad a \in [0,1]. \tag{4.11}$$

Proposition 4.2. *In Example* 4.2, *the following assertions* (i)–(iv) *hold*, *where the policy* $f_a$ *is defined in* (4.8):

(i) *The expected total reward at state* 1 *under policy* $f_a$ *is determined by*

$$\lambda V_\lambda(1, f_a) = \begin{cases} \log\left(\dfrac{1-a}{e^{(1-a)/(1-a/2)} - a}\right) & \text{if } a \in [0,1), \\ 0 & \text{if } a = 1. \end{cases}$$

(ii) *The mapping* $a \mapsto V_\lambda(f_a, 1)$ *is increasing in* $a \in [0,1)$.

(iii) *As* $a \nearrow 1$, $V_\lambda(f_a, 1) \nearrow L$, *where* $L = \lambda^{-1}\log(1/2) = |\lambda|^{-1}\log(2)$.

*Moreover*,

(iv) $L = V_\lambda^*(1)$ *and thus, no stationary policy is* $\lambda$*-optimal.*

P r o o f. (i) As in the proof of Proposition 4.1, it follows that $V_\lambda(f_1, 1) = 0$, and for $a \in [0,1)$,

$$\lambda V_\lambda(f_a, 1) = \log\left(\frac{e^{\lambda r(a)}(1-a)}{1 - ae^{\lambda r(a)}}\right) = \log\left(\frac{1-a}{e^{-\lambda r(a)} - a}\right) \tag{4.12}$$

now the conclusion follows since, in the present case, $-\lambda r(a) = |\lambda| r(a) = (1-a)/(1-a/2)$; see (4.11) and recall that $\lambda < 0$.

(ii) Straightforward calculations using (4.12) yield that for $a \in [0,1)$,

$$\lambda \frac{dV_\lambda(f_a, 1)}{da} = \frac{e^{-\lambda r(a)}}{(1-a)(e^{-\lambda r(a)} - a)}[e^{\lambda r(a)} - 1 + \lambda r'(a)(1-a)], \tag{4.13}$$

where the derivative is from the right at $a = 0$. To continue, notice that, since $\lambda$ is negative, (4.11) yields that $\lambda r(a) = -(1-a)/(1-a/2)$, so that $\lambda r'(a) = 1/(1-a/2) - (1-a)/[2(1-a/2)^2]$, and then $\lambda r'(a)(1-a) = (1-a)/(1-a/2) - (1-a)^2/[2(1-a/2)^2]$, that is, $\lambda r'(a)(1-a) = -\lambda r(a) - (\lambda r(a))^2/2$, so that

$$e^{\lambda r(a)} - 1 + \lambda r'(a)(1-a) = e^{\lambda r(a)} - 1 - \lambda r(a) - \frac{(\lambda r(a))^2}{2}.$$

Set $G(x) = e^x - 1 - x - x^2/2$ and observe that, by the strict convexity of the exponential function, $G'(x) = e^x - 1 - x > 0$ for every $x \neq 0$. Thus, $G(x) < G(0) = 0$ if $x < 0$, and if we use this fact with $x = \lambda r(a)$, the last displayed equality implies that

$$e^{\lambda r(a)} - 1 + \lambda r'(a)(1-a) < 0, \quad a \in [0,1).$$

Since the quotient in (4.13) is positive, it follows that $\lambda dV_\lambda(f_a, 1)/da < 0$, and then, since $\lambda$ is negative, $dV_\lambda(f_a, 1)/da$ is positive for $a \in [0,1)$. This establishes part (ii) and the other parts can be obtained along the same lines as in the proof of Proposition 4.1. ∎

In Example 4.2 the continuity-compactness conditions of Remark 3.1 are satisfied, so that there exists a policy $f \in \mathbb{F}$ such that $f(x)$ maximizes the right-hand side of the $\lambda$-OE. By Proposition 4.2, that policy is not $\lambda$-optimal and, from Theorem 4.1, the Markov chain induced by $f$ does not have the

unichain property. In fact, it can be directly verified that the unique policy satisfying (4.1) is the policy $f_1$ (see (4.8)) which, as already noted, does not have the unichain property. To summarize, the two previous examples have shown that, regardless of the sign of the nonzero risk-sensitivity coefficient, when a policy $f$ satisfying (4.1) does not have the unichain property, the existence of a $\lambda$-optimal stationary policy cannot be ensured. After this discussion, attention is now turned to the proof of Theorem 4.1.

**5. Technical preliminaries.** This section contains some preliminary facts that will be used to establish Theorem 4.1 in the next section. The starting point is the following lemma concerning the asymptotic behaviour of the expected utility of $V_\lambda^*(X_n)$.

LEMMA 5.1. *Suppose that Assumptions* 2.1 *and* 2.2 *hold true, and that the reward function is nonnegative. Let $f \in \mathbb{F}$ be a policy satisfying* (4.1) *and assume that $f$ has the unichain property described in the statement of Theorem* 4.1. *For each $x \in S$ and $n \in \mathbb{N}$, let $L_n(x)$ be the certain equivalent of $V_\lambda^*(X_n)$ with respect to $U_\lambda(\cdot)$, so that $L_n(x)$ is determined by*

$$U_\lambda(L_n(x)) = E_f[U_\lambda(V_\lambda^*(X_n)) \,|\, X_0 = x].$$

*In this case, there exists a nonnegative constant $C$ such that*

$$\lim_{n\to\infty} L_n(x) = C, \quad x \in S.$$

*Moreover,*

$$V_\lambda^*(x) \geq C, \quad x \in S.$$

P r o o f. Since $f$ satisfies (4.1), the Markov property yields

$$U_\lambda(V_\lambda^*(X_n)) = E_f[e^{\lambda R(X_n,A_n)} U_\lambda(V_\lambda^*(X_{n+1})) \,|\, X_n], \tag{5.1}$$

and, from the definition of the functions $L_k(\cdot)$, it follows that

$$E_f[U_\lambda(V_\lambda^*(X_{k+1})) \,|\, X_1 = y] = E_f[U_\lambda(V_\lambda^*(X_k)) \,|\, X_0 = y] = U_\lambda(L_k(y)),$$

so that $E_f[U_\lambda(V_\lambda^*(X_{k+1})) \,|\, X_1] = U_\lambda(L_k(X_1))$, and then

$$\begin{aligned} U_\lambda(L_{k+1}(x)) &= E_f[U_\lambda(V_\lambda^*(X_{k+1})) \,|\, X_0 = x] \\ &= E_f[E_f[U_\lambda(V_\lambda^*(X_{k+1})) \,|\, X_1] \,|\, X_0 = x] \\ &= E_f[U_\lambda(L_k(X_1)) \,|\, X_0 = x]. \end{aligned} \tag{5.2}$$

It will be verified that

$$L_n(\cdot) \geq L_{n+1}(\cdot) \geq 0. \tag{5.3}$$

To establish this assertion notice that, since $R \geq 0$, the optimal value function is nonnegative, by Remark 2.1(iii), and thus $L_n(\cdot) \geq 0$ for every $n \in \mathbb{N}$. Next, consider the risk-seeking and risk-averse cases:

CASE 1: $\lambda > 0$. In this case $U_\lambda(\cdot) \geq 0$, and since $R \geq 0$, it follows that $e^{\lambda R(\cdot,\cdot)} \geq 1$, so that (5.1) implies that $U_\lambda(V_\lambda^*(X_n)) \geq E_f[U_\lambda(V_\lambda^*(X_{n+1})) \,|\, X_n]$ for every $n \in \mathbb{N}$; consequently, for every $x \in S$,

$$\begin{aligned} U_\lambda(L_n(x)) &= E_f[U_\lambda(V_\lambda^*(X_n)) \,|\, X_0 = x] \\ &\geq E_f[E_f[U_\lambda(V_\lambda^*(X_{n+1})) \,|\, X_n] \,|\, X_0 = x] \\ &= E_f[U_\lambda(V_\lambda^*(X_{n+1})) \,|\, X_0 = x] = U_\lambda(L_{n+1}(x)) \end{aligned}$$

and as $U_\lambda(\cdot)$ is increasing, it follows that $L_n(\cdot) \geq L_{n+1}(\cdot)$, since $x \in S$ was arbitrary.

CASE 2: $\lambda < 0$. Observe that $R \geq 0$ implies that $e^{\lambda R(\cdot,\cdot)} \leq 1$. Since $U_\lambda(\cdot) \leq 0$, this yields $e^{\lambda R(X_n,A_n)} U_\lambda(V_\lambda^*(X_{n+1})) \geq U_\lambda(V_\lambda(X_{n+1}))$, and thus (5.1) yields

$$\begin{aligned} U_\lambda(V_\lambda^*(X_n)) &= E_f[e^{\lambda R(X_n,A_n)} U_\lambda(V_\lambda^*(X_{n+1})) \,|\, X_n] \\ &\geq E_f[U_\lambda(V_\lambda^*(X_{n+1})) \,|\, X_n]; \end{aligned}$$

therefore, taking expectation with respect to $P_f[\cdot \,|\, X_0 = x]$ shows that

$$\begin{aligned} U_\lambda(L_n(x)) &= E_f[U_\lambda(V_\lambda^*(X_n)) \,|\, X_0 = x] \\ &\geq E_f[U_\lambda(V_\lambda^*(X_{n+1})) \,|\, X_0 = x] \\ &= U_\lambda(L_{n+1}(x)) \end{aligned}$$

and thus (5.3) is also valid in the risk-averse case. The proof now goes as follows: First notice that $E_f[U_\lambda(V_\lambda^*(X_0)) \,|\, X_0 = x] = U_\lambda(V_\lambda^*(x))$, so that

$$L_0(\cdot) = V_\lambda^*(\cdot). \tag{5.4}$$

From (5.3), it follows that there exists a function $L : S \to [0, \infty)$ such that

$$\lim_{n\to\infty} L_n(x) = L(x), \quad x \in S. \tag{5.5}$$

Taking the limit as $k \nearrow \infty$ in (5.2) and using the bounded convergence theorem implies

$$U_\lambda(L(x)) = E_f[U_\lambda(L(X_1)) \,|\, X_0 = x], \quad x \in S,$$

and, via the Markov property, an induction argument yields that

$$U_\lambda(L(x)) = E_f[U_\lambda(L(X_n)) \,|\, X_0 = x].$$

Therefore, for every $x \in S$,

$$U_\lambda(L(x)) = \lim_{n\to\infty} \frac{1}{n+1} \sum_{t=0}^{n} E_f[U_\lambda(L(X_t)) \,|\, X_0 = x] = \sum_y \mu(y) U_\lambda(y),$$

where $\mu(\cdot)$ is the unique invariant distribution of the Markov chain induced by $f$; see, for instance, Loève (1977). Thus, $U_\lambda(L(\cdot))$ is constant, and since

$U_\lambda(\cdot)$ is strictly monotone, it follows that $L(\cdot) \equiv C$ for some constant $C$. To conclude, observe that combining the convergence in (5.5) with (5.3) and (5.4), it follows that $V_\lambda^*(\cdot) = L_0(\cdot) \geq C \geq 0$. ∎

LEMMA 5.2. *Under the assumptions of Lemma* 5.1, *for each* $x \in S$,

$$V_\lambda^*(X_n) \xrightarrow{P_f[\cdot \,|\, X_0 = x]} C,$$

*i.e., for each* $\varepsilon > 0$, $P_f[|V_\lambda^*(X_n) - C| > \varepsilon \,|\, X_0 = x] \to 0$ *as* $n \to \infty$.

P r o o f. First, it will be verified that

$$1 = \lim_{n\to\infty} E_f[e^{|\lambda(V_\lambda(X_n)-C)|}]. \tag{5.6}$$

To prove this convergence, notice that by Lemma 5.1,

$$U_\lambda(C) = \lim_{n\to\infty} U_\lambda(L_n(x)) = \lim_{n\to\infty} E_f[U_\lambda(V_\lambda^*(X_n))],$$

which, by the definition of $U_\lambda(\cdot)$ in (2.1), is equivalent to

$$1 = \lim_{n\to\infty} E_f[e^{\lambda(V_\lambda^*(X_n)-C)}].$$

Since $V_\lambda^*(\cdot) - C \geq 0$, by Lemma 5.1, this relation yields (5.6) when $\lambda > 0$. Suppose now $\lambda$ is negative, and observe that in this case the above convergence can be written as $1 = \lim_{n\to\infty} E_f[1/e^{|\lambda|(V_\lambda^*(X_n)-C)}]$, which is equivalent to

$$0 = \lim_{n\to\infty} E_f\left[\frac{e^{|\lambda|(V_\lambda^*(X_n)-C)} - 1}{e^{|\lambda|(V_\lambda^*(X_n)-C)}}\right]. \tag{5.7}$$

Setting $M = \max_{x\in S}\{V_\lambda^*(x) - C\}$, and recalling that $V_\lambda^*(\cdot) - C$ is nonnegative, we obtain

$$\frac{e^{|\lambda|(V_\lambda^*(X_n)-C)} - 1}{e^{|\lambda|(V_\lambda^*(X_n)-C)}} \geq \frac{e^{|\lambda|(V_\lambda^*(X_n)-C)} - 1}{e^{|\lambda|M}} = \frac{e^{|\lambda(V_\lambda^*(X_n)-C)|} - 1}{e^{|\lambda|M}} \geq 0,$$

so that (5.7) implies that (5.6) is also valid for $\lambda < 0$. To conclude, take logarithms on both sides of (5.6) to obtain

$$0 = \log(\lim_{n\to\infty} E_f[e^{|\lambda(V_\lambda^*(X_n)-C)|}]) = \lim_{n\to\infty} \log(E_f[e^{|\lambda(V_\lambda^*(X_n)-C)|}]);$$

since the logarithmic function is strictly concave, Jensen's inequality yields

$$\log(E_f[e^{|\lambda(V_\lambda^*(X_n)-C)|}]) \geq E_f[|\lambda(V_\lambda^*(X_n) - C)|]$$

and the above convergence implies that $0 = \lim_{n\to\infty} E_f[|\lambda(V_\lambda^*(X_n) - C)|]$; since $\lambda \neq 0$, the conclusion now follows from Markov's inequality. ∎

**6. Proof of the main result.** After the preliminaries in the previous section, Theorem 5.1 can be established as follows.

*Proof of Theorem 5.1.* Let $f$ be a stationary policy satisfying (4.1) and assume that $f$ has a unique positive recurrent class. To begin with, it will be shown that the nonnegative constant $C$ in Lemmas 5.1 and 5.2 is zero:

$$C = 0. \tag{6.1}$$

To establish this equality, notice that (2.2) implies that

$$\begin{aligned} U_\lambda(V_\lambda^*(x) - C) &= e^{-\lambda C} U_\lambda(V_\lambda^*(x)) \\ &= e^{-\lambda C} \sup_{a \in A(x)} \Big[ e^{\lambda R(x,a)} \sum_y p_{xy}(a) U_\lambda(V_\lambda^*(y)) \Big] \\ &= \sup_{a \in A(x)} \Big[ e^{\lambda R(x,a)} \sum_y p_{xy}(a) e^{-\lambda C} U_\lambda(V_\lambda^*(y)) \Big] \\ &= \sup_{a \in A(x)} \Big[ e^{\lambda R(x,a)} \sum_y p_{xy}(a) U_\lambda(V_\lambda^*(y) - C) \Big] \end{aligned}$$

so that $V_\lambda^*(\cdot) - C$ satisfies the $\lambda$-OE. Since $V_\lambda^*(\cdot) - C \geq 0$ by Lemma 5.1, it follows from Lemma 4.1 that $V_\lambda^*(\cdot) - C \geq V_\lambda^*(\cdot)$, and thus $C \leq 0$; consequently, (6.1) holds, since $C$ is nonnegative. Next, observe that an induction argument using the Markov property yields that for every $x \in S$ and $n \in \mathbb{N}$,

$$U_\lambda(V_\lambda^*(x)) = E_f[e^{\lambda \sum_{t=0}^n R(X_t, A_t)} U_\lambda(V_\lambda^*(X_{n+1})) \,|\, X_0 = x]. \tag{6.2}$$

Observe now the following facts (a)–(c):

(a) Since $R \geq 0$, $e^{\lambda \sum_{t=0}^n R(X_t, A_t)}$ lies between $\min\{1, e^{\lambda R(X_0, A_0)}\}$ and $\max\{1, e^{\lambda \sum_{t=0}^\infty R(X_t, A_t)}\}$.

(b) $e^{\lambda \sum_{t=0}^n R(X_t, A_t)}$ converges to $e^{\lambda \sum_{t=0}^\infty R(X_t, A_t)}$ everywhere.

(c) With respect to $P_f[\cdot \,|\, X_0 = x]$, $U_\lambda(V_\lambda^*(X_n))$ converges to $U_\lambda(0)$ in probability; this property follows from the continuity of $U_\lambda(\cdot)$ together with Lemma 5.2 and (6.1).

From (b) and (c), it follows that, for arbitrary $x \in S$, the following convergence holds in $P_f[\cdot \,|\, X_0 = x]$-measure:

$$e^{\lambda \sum_{t=0}^n R(X_t, A_t)} U_\lambda(V_\lambda^*(X_{n+1})) \to e^{\lambda \sum_{t=0}^\infty R(X_t, A_t)} U_\lambda(0).$$

Since $E_f[e^{\lambda \sum_{t=0}^\infty R(X_t, A_t)}] = \operatorname{sign}(\lambda) U_\lambda(V_\lambda(f, x)) < \infty$, by Assumption 2.2, and $|U_\lambda(V_\lambda^*(X_{n+1}))| \leq e^{|\lambda| M}$, where $M = \max_{x \in S}\{|V_\lambda^*(x)|\}$, the above convergence and (a) together imply, via the dominated convergence theorem, that

$$\begin{aligned} \lim_{n \to \infty} E_f[e^{\lambda \sum_{t=0}^n R(X_t, A_t)} U_\lambda(V_\lambda^*(X_{n+1})) \,|\, X_0 = x] \\ = E_f[e^{\lambda \sum_{t=0}^\infty R(X_t, A_t)} U_\lambda(0) \,|\, X_0 = x]. \end{aligned}$$

Combining this convergence with (6.2) shows that

$$\begin{aligned} U_\lambda(V_\lambda^*(x)) &= E_f[e^{\lambda \sum_{t=0}^{\infty} R(X_t,A_t)} U_\lambda(0)) \,|\, X_0 = x] \\ &= E_f[U_\lambda(e^{\lambda \sum_{t=0}^{\infty} R(X_t,A_t)}) \,|\, X_0 = x], \end{aligned}$$

where the second equality follows from (2.2). Then (2.6) yields $U_\lambda(V_\lambda^*(x)) = U_\lambda(V_\lambda(f,x))$, and since $U_\lambda(\cdot)$ is increasing, it follows that $V_\lambda^*(x) = V_\lambda(f,x)$, establishing the $\lambda$-optimality of the policy $f$, since $x \in S$ was arbitrary. ∎

**7. Conclusion.** This work considered finite-state MDP's endowed with the risk-sensitive expected total-reward criterion given in (2.5)–(2.7). The $\lambda$-optimality of a stationary policy $f$ achieving the maximum in the $\lambda$-OE was established whenever $f$ has the unichain property, and examples were given to show that, regardless of the sign of the risk-sensitivity coefficient, when the unichain property fails the existence of an optimal stationary policy cannot be generally ensured; as already mentioned, these results provide an answer to a question in Puterman (1994) when it is interpreted in the risk-sensitive context. The arguments in the paper are concentrated on stationary policies, but at this point it is convenient to mention that, in Examples 4.1 and 4.2, it is not difficult to verify that a $\lambda$-optimal policy does not exist even within the class $\mathcal{P}$ of *all* policies.

## References


M. G. Ávila-Godoy (1998), *Controlled Markov chains with exponential risk-sensitive criteria*: *modularity*, *structured policies and applications*, Ph.D. Dissertation, Dept. of Math., Univ. of Arizona, Tucson, AZ.

R. Cavazos-Cadena and E. Fernández-Gaucherand (1999), *Controlled Markov chains with risk-sensitive criteria*: *average cost*, *optimality equations*, *and optimal solutions*, Math. Methods Oper. Res. 43, 121–139.

R. Cavazos-Cadena and R. Montes-de-Oca (1999), *Optimal stationary policies in controlled Markov chains with the expected total-reward criterion*, Research Report No. 1.01.010.99, Univ. Autónoma Metropolitana, Campus Iztapalapa, México, D.F.

P. C. Fishburn (1970), *Utility Theory for Decision Making*, Wiley, New York.

W. H. Fleming and D. Hernández-Hernández (1997), *Risk-sensitive control of finite machines on an infinite horizon I*, SIAM J. Control Optim. 35, 1790–1810.

O. Hernández-Lerma (1989), *Adaptive Markov Control Processes*, Springer, New York.

K. Hinderer (1970), *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Oper. Res. 33, Springer, New York.

R. A. Howard and J. E. Matheson (1972), *Risk-sensitive Markov decision processes*, Management Sci. 18, 356–369.

M. Loève (1977), *Probability Theory I*, 4th ed., Springer, New York.

J. W. Pratt (1964), *Risk aversion in the small and in the large*, Econometrica 32, 122–136.
M. L. Puterman (1994), *Markov Decision Processes*, Wiley, New York.
S. M. Ross (1970), *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco.
R. Strauch (1966), *Negative dynamic programming*, Ann. Math. Statist. 37, 871–890.

Rolando Cavazos-Cadena
Departamento de Estadística y Cálculo
Universidad Autónoma Agraria Antonio Narro
Buenavista, Saltillo COAH 25315, México
E-mail: rcavazos@narro.uaaan.mx

Raúl Montes-de-Oca
Departamento de Matemáticas
Universidad Autónoma Metropolitana
Campus Iztapalapa
Avenida Michoacán y La Purísima s/n
Col. Vicentina
México, D.F. 09340, México
E-mail: momr@xanum.uam.mx

*Received on 20.4.1999;*
*revised version on 5.10.1999*