APPLICATIONES MATHEMATICAE
28,1 (2001), pp. 93–109

Rolando Cavazos-Cadena (Saltillo)
Raúl Montes-de-Oca (México)

STATIONARY OPTIMAL POLICIES IN A CLASS OF MULTICHAIN POSITIVE DYNAMIC PROGRAMS WITH FINITE STATE SPACE AND RISK-SENSITIVE CRITERION

Abstract. This work concerns Markov decision processes with finite state space and compact action sets. The decision maker is supposed to have a constant-risk sensitivity coefficient, and a control policy is graded via the risk-sensitive expected total-reward criterion associated with nonnegative one-step rewards. Assuming that the optimal value function is finite, under mild continuity and compactness restrictions the following result is established: If the number of ergodic classes when a stationary policy is used to drive the system depends continuously on the policy employed, then there exists an optimal stationary policy, extending results obtained by Schäl (1984) for *risk-neutral* dynamic programming. We use results recently established for *unichain* systems, and analyze the general multichain case via a reduction to a model with the unichain property.

1. Introduction. This work deals with Markov decision processes (MDPs) with finite state space, compact action sets and nonnegative rewards that, together with the transition law, are continuous functions of the actions. The decision maker is assumed to have a constant (and nonnull) risk-sensitivity coefficient, which leads to evaluate random rewards via an expected value involving an exponential utility function (Pratt (1964), Fishburn (1970)), and the perfomance of a decision policy is measured by the corresponding *risk-sensitive* expected *total-reward* criterion. Although the Markov chain associated with a stationary policy can have several min-

2000 *Mathematics Subject Classification*: 93E20, 90C40.
Key words and phrases: utility function, structural stability condition, closed sets, ergodic class, invariant distribution.
This work was generously supported by the PSF Organization under Grant No. 30-250-99-05.

imal closed sets (the multichain case), their number is supposed to depend continuously on the policy employed. Under this latter structural stability condition (Schweitzer (1968)), the main result of this note, stated as Theorem 3.1 below, establishes the existence of an optimal stationary policy.

The result in this work extends theorems in Schäl (1984, 1986), where optimal stationary policies were obtained for risk-neutral multichain dynamic programs via the discounted criterion, and in Cavazos-Cadena and Montes-de-Oca (2000a), where risk-sensitive *unichain* MDPs were analyzed. The strategy to prove Theorem 3.1 does not uses the discounted criterion, but is based on the construction of a new MDP satisfying the unichain property, so that, essentially, the main result of this note is obtained from Theorem 4.1 in the last mentioned work.

The organization of the present paper is as follows: After a short description of the decision model, Section 2 contains the basic notions and the properties of the optimal value function that will be used later on. Next, in Section 3 the result on the existence of optimal stationary policies for multichain MDPs is stated as Theorem 3.1, and the strategy to establish this result is described. The necessary preliminaries to follow the outlined route are established in Section 4, and the proof of Theorem 3.1 is given in Section 5 before the concluding remarks in Section 6.

2. Decision model. Throughout the remainder of the paper $M = (S, A, \{A(x)\}, R, P)$ denotes the usual MDP, where the state space S is assumed to be a *finite* set endowed with the discrete topology, the metric space A is the control (or action) set and, for each $x \in S$, $A(x) \subset A$ is the nonempty subset of admissible actions at state x. On the other hand, $R(\cdot, \cdot)$ is the reward function defined on the class $\mathbb{K} := \{(x, a) \mid a \in A(x),\ x \in S\}$ of *admissible pairs*, and $P = [p_{x\,y}(\cdot)]$ is the controlled transition law. The interpretation of M is as follows: At each time $t \in \mathbb{N} := \{0, 1, \ldots\}$ 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 $p_{x\,y}(a)$; this is the Markov property of the decision model.

Assumption 2.1. (i) For each $x \in S$, the action set $A(x)$ is a compact subspace of A.

(ii) For every $x, y \in S$, the mapping $a \mapsto p_{x\,y}(a)$ is continuous on $A(x)$.

(iii) The reward function is nonnegative: $R(x, a) \geq 0$, $(x, a) \in \mathbb{K}$.

(iv) For each $x \in S$, $a \mapsto R(x, a)$ is a continuous function on $A(x)$.

Policies. A policy is a (measurable) rule for choosing actions which, at each time $t \in \mathbb{N}$, may depend on the current state as well as on the record

of previous states and controls. For 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. Set $\mathbb{F} := \prod_{x \in S} A(x)$, so that $\mathbb{F}$ consists of all (choice) functions $f : S \to A$ satisfying that $f(x) \in A(x)$ for every $x \in S$; since each set $A(x)$ is a compact subset of the metric space A, $\mathbb{F}$ itself is a compact metric space in the product topology (Dugundji (1966)). A policy $\pi \in \mathcal{P}$ is stationary if there exists $f \in \mathbb{F}$ such that the action prescribed by π when $X_t = x$ is observed is *always* $f(x)$; the class of stationary policies is naturally identified with $\mathbb{F}$.

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)), and the following terminology will be used.

Definition 2.1. Let $f \in \mathbb{F}$ be fixed.

(i) A nonempty set $C \subset S$ is *f-closed* if $\sum_{y \in C} p_{x\,y}(f(x)) = 1$ for every $x \in C$.

(ii) $C^* \subset S$ is a *minimal f-closed set* if

(a) C^* is f-closed, and

(b) if an f-closed set C satisfies $C \subset C^*$, then $C = C^*$.

(iii) The function $\mathcal{E} : \mathbb{F} \to \mathbb{N}$ is given by

$$\mathcal{E}(f) = \text{number of minimal } f\text{-closed sets.}$$

(iv) The decison model M is *unichain* if $\mathcal{E}(f) = 1$ for every $f \in \mathbb{F}$, whereas M is *multichain* when $\mathcal{E}(\cdot)$ is not identically one.

Remark 2.1 (Section 8 of Billingsley (1995), Chapter 3 of Loève (1977)).

(i) Since the state space is finite, for each $f \in \mathbb{F}$ the class of minimal f-closed sets coincides with the family of ergodic classes of the Markov chain associated with f. More precisely, let $C_1^*, \ldots, C_k^*$ be the minimal f-closed sets and denote by $\mathcal{R}(f)$ the set of all (necessarily positive) recurrent states with respect to the Markov chain induced by f. With this notation, statements (a)–(c) below are valid:

(a) $P_f[X_n = y \text{ for some } n \in \mathbb{N} \setminus \{0\} \,|\, X_0 = x] = 1$ for all $x, y \in C_i^*$ and $i = 1, \ldots, k$,

(b) for each $x \in C_i^*$ and $i = 1, \ldots, k$,

$$P_f[X_n \in C_i^* \text{ for all } n \in \mathbb{N} \,|\, X_0 = x] = 1,$$

(c) $C_i^* \cap C_j^* = \emptyset$ if $i \neq j$, and
(d) $\mathcal{R}(f) = C_1^* \cup \ldots \cup C_k^*$.

(ii) For a minimal f-closed set C^*, there exists a probability distribution $\mu : S \to [0,1]$ such that $\mu(y) = 0$ for $y \in S \setminus C^*$, and

$$\mu(y) = \sum_x \mu(x) p_{x\,y}(f(x)), \quad y \in S, \tag{2.1}$$

i.e., μ is an invariant distribution of the Markov chain induced by f.

Conversely,

(iii) Suppose that the probability distribution $\mu : S \to [0,1]$ satisfies (2.1). Then the support of μ, defined by $\mathrm{supp}(\mu) = \{x \in S \mid \mu(x) \neq 0\}$, contains a minimal f-closed set.

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

$$U_\lambda(x) := \begin{cases} \mathrm{sign}(\lambda) e^{\lambda x} & \text{if } \lambda \neq 0, \\ x & \text{when } \lambda = 0; \end{cases} \tag{2.2}$$

notice that

$$U_\lambda(x + c) = e^{\lambda c} U_\lambda(x), \quad x, c \in \mathbb{R},\ \lambda \neq 0. \tag{2.3}$$

A controller with risk-sensitivity λ grades a random reward Y via the expectation of $U_\lambda(Y)$. If the initial state is $X_0 = x \in S$, and policy π is used to drive the system, the expected utility of the total reward earned at times $t \in \mathbb{N}$ is $E_x^\pi[U_\lambda(\sum_{t=0}^\infty R(X_t, A_t))]$. The *$\lambda$-sensitive expected total reward* at x under policy π, denoted by $V_\lambda(\pi, x)$, is implicitly determined by

$$U_\lambda(V_\lambda(\pi, x)) = E_\pi\Big[U_\lambda\Big(\sum_{t=0}^\infty R(X_t, A_t)\Big) \,\Big|\, X_0 = x\Big], \tag{2.4}$$

an expression that, for $\lambda \neq 0$, is equivalent to

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

see (2.2) and notice that, since the reward function is nonnegative, the expectations in the above expressions are well defined and the inequality $0 \leq V_\lambda(\pi, x)$ is always valid. The *λ-optimal value function* is

$$V_\lambda^*(x) = \sup_\pi V_\lambda(\pi, x), \quad x \in S, \tag{2.6}$$

and a policy π^* is *λ-optimal* if $V_\lambda(\pi^*, x) = V_\lambda^*(x)$ for all $x \in S$.

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

Remark 2.2. When $\lambda > 0$, the utility function $U_\lambda(\cdot)$ is convex, and via Jensen's inequality, (2.2) and (2.4) yield that $V_\lambda(\pi, x) \geq V_0(\pi, x)$; similarly,

$V_\lambda(\pi, x) \le V_0(\pi, x)$ if $\lambda < 0$. A decision maker grading a random reward Y according to the expectation of $U_\lambda(Y)$ is referred to as *risk-seeking* if $\lambda > 0$, or *risk-averse* when $\lambda < 0$; if $\lambda = 0$, the controller is *risk-neutral*.

Risk-sensitive optimality equation. Throughout the remainder the risk-sensitivity coefficient λ is supposed to be nonnull. In this case, under Assumptions 2.1 and 2.2, $V_\lambda^*(\cdot)$ in (2.6) satisfies the following λ-*optimality equation* (λ-OE):

$$U_\lambda(V_\lambda^*(x)) = \sup_{a \in A(x)} \Big[e^{\lambda R(x,a)} \sum_y p_{x\,y}(a) U_\lambda(V_\lambda^*(y)) \Big], \quad x \in S; \tag{2.7}$$

see, for instance, Ávila-Godoy (1998), or Cavazos-Cadena and Montes-de-Oca (2000a). Moreover, since the state space is finite, the term within brackets in this equality is a continuous function of $a \in A(x)$, and the compactness of action sets immediately yields that there exists a policy $f \in \mathbb{F}$ such that, for each $x \in S$, $f(x) \in A(x)$ maximizes the right-hand side of the λ-OE.

LEMMA 2.1. *Suppose that Assumptions* 2.1 *and* 2.2 *hold, and let the stationary policy* $f \in \mathbb{F}$ *be such that*

$$U_\lambda(V_\lambda^*(x)) = e^{\lambda R(x, f(x))} \sum_y p_{x\,y}(f(x)) U_\lambda(V_\lambda^*(y)), \quad x \in S. \tag{2.8}$$

If, additionally, $\mathcal{E}(f) = 1$, *then* f *is* λ*-optimal.*

According to this lemma, a stationary policy f obtained by maximizing the right-hand side of (2.7) is λ-optimal whenever f induces a Markov chain with a *single* ergodic class. A proof can be found in Cavazos-Cadena and Montes-de-Oca (2000a), where it was also shown that the λ-optimality of the policy f in (2.8) cannot be generally ensured when $\mathcal{E}(f) \neq 1$. To analyze the existence of optimal stationary policies in the general multichain case, the following characterization of $V_\lambda^*(\cdot)$ as the minimal solution of the λ-OE will be useful.

LEMMA 2.2 (Cavazos-Cadena and Montes-de-Oca (2000a)). *Given* $\lambda \neq 0$, *suppose that the model* M *satisfies Assumption* 2.1 *and let* $W : S \to [0, \infty)$ *be such that*

$$U_\lambda(W(x)) \ge \sup_{a \in A(x)} \Big[e^{\lambda R(x,a)} \sum_{y \in S} p_{x\,y}(a) U_\lambda(W(y)) \Big], \quad x \in S.$$

In this case $W(\cdot) \ge V_\lambda^*(\cdot)$.

Finally, for every subset G of the state space, define the first positive arrival time to G by

$$T_G = \min\{n > 0 \mid X_n \in G\}, \tag{2.9}$$

where, as usual, the minimum of the empty set is ∞.

3. Main result. As already mentioned, the main purpose of this note is to analyze the existence of λ-optimal stationary policies for *multichain* MDPs. In addition to Assumptions 2.1 and 2.2, this problem will be studied under the following requirement, which is referred to as the *structural stability condition* (Schweitzer (1968), Schäl (1986), Cavazos-Cadena, Feinberg and Montes-de-Oca (2000)).

ASSUMPTION 3.1. The function $\mathcal{E}(\cdot)$ is continuous.

REMARK 3.1. (i) Under Assumption 2.1(ii), it is not difficult to see that $\mathcal{E}(\cdot)$ is always an upper semicontinuous function, i.e., if $\{f_n\} \subset \mathbb{F}$ is such that $\lim_n f_n = f \in \mathbb{F}$, then $\limsup_n \mathcal{E}(f_n) \leq \mathcal{E}(f)$ (Schäl (1984)). However, strict inequality can occur. For instance, in Examples 4.1 and 4.2 in Cavazos-Cadena and Montes-de-Oca (2000a), the stationary policies can be indexed by the set $[0,1]$, and $\lim_{a\downarrow 0} f_a = f_0$, but $\mathcal{E}(f_0) = 2 > 1 = \mathcal{E}(f_a)$ for every $a \in (0,1]$, so that an ergodic class "suddenly" appears when approaching policy f_0; such a phenomenon cannot occur under Assumption 3.1. As a consequence of Theorem 3.1 below, in the two examples mentioned above the lack of λ-optimal stationary policies can be traced back to the discontinuity of the function $\mathcal{E}(\cdot)$.

(ii) Since $\mathcal{E}(\cdot)$ is integer-valued, under Assumption 3.1 this function must be constant in each connected component of the space $\mathbb{F}$. In particular, if all the action sets are connected, so is $\mathbb{F}$ and $\mathcal{E}(\cdot)$ must be constant on its whole domain when Assumption 3.1 holds (Dugundji (1966)).

The following is the main result of this paper.

THEOREM 3.1. *Let $\lambda \neq 0$ be fixed, and suppose that Assumptions* 2.1, 2.2 *and* 3.1 *hold. In this case, there exists a λ-optimal stationary policy for the model M.*

This theorem extends results that, for risk-neutral dynamic programming, were obtained by Schäl (1984) via the *discounted criterion*; see also Cavazos-Cadena, Feinberg and Montes-de-Oca (2000). Roughly, in the present risk-sensitive context, Theorem 3.1 will be derived, after performing an appropriate reduction to the unichain case, from Lemma 2.1. The starting point in this route is Theorem 3.2 below, whose statement involves the following notation.

DEFINITION 3.1. Suppose that the MDP $M = (S, A, \{A(x)\}, R, P)$ satisfies Assumptions 2.1 and 2.2.

(i) The kernel $\mathcal{K}$ of the optimal value function $V_\lambda^*(\cdot)$ is given by

$$\mathcal{K} = \{x \in S \mid V_\lambda^*(x) = 0\}.$$

(ii) Given an object Δ outside of S and a fixed action $a_\Delta \in A$, define a

new MDP

$$\widetilde{M} = (\widetilde{S}, A, \{\widetilde{A}(x)\}, \widetilde{R}, \widetilde{P})$$

as follows:

$$\widetilde{S} = \{\Delta\} \cup (S \setminus \mathcal{K}), \quad \widetilde{A}(x) = A(x), \quad x \in S \setminus \mathcal{K}, \quad \widetilde{A}(\Delta) = \{a_\Delta\};$$

the reward function $\widetilde{R}$ is defined by

$$\widetilde{R}(x,a) = R(x,a), \quad x \in S \setminus \mathcal{K},\ a \in A(x), \quad \widetilde{R}(\Delta, a_\Delta) = 0,$$

whereas the transition law $\widetilde{P} = [\widetilde{p}_{x\,y}(\cdot)]$ is determined as follows:

$$\widetilde{p}_{\Delta\,\Delta}(a_\Delta) = 1,$$

and for $x \in S \setminus \mathcal{K}$,

$$\widetilde{p}_{x\,y}(a) = \begin{cases} p_{x\,y}(a) & \text{if } y \in S \setminus \mathcal{K}, \\ \sum_{z \in \mathcal{K}} p_{x\,z}(a) & \text{if } y = \Delta. \end{cases}$$

(iii) The optimal value function and the class of all policies for the model $\widetilde{M}$ are denoted by $\widetilde{V}^*_\lambda(\cdot)$ and $\widetilde{\mathcal{P}}$, respectively.

REMARK 3.2. (i) Let $x \in \mathcal{K}$ and $a \in A(x)$ be fixed. From the λ-OE in (2.7) it follows that

$$U_\lambda(0) = U_\lambda(V^*_\lambda(x)) \geq e^{\lambda R(x,a)} \sum_{y \in S} p_{x\,y}(a) U_\lambda(V^*_\lambda(y)) = \sum_{y \in S} p_{x\,y}(a) U_\lambda(R(x,a) + V^*_\lambda(y)), \tag{3.1}$$

where the equality is due to (2.3). Since $R(\cdot,\cdot)$ and $V^*_\lambda(\cdot)$ are nonnegative and the utility function is strictly increasing, (3.1) yields that $U_\lambda(0) \geq \sum_{y \in S} p_{x\,y}(a) U_\lambda(R(x,a) + V^*_\lambda(y)) \geq U_\lambda(0)$, and consequently, $R(x,a) + V^*_\lambda(y) = 0$ if $p_{x\,y}(a) > 0$. This discussion can be summarized as follows:

$$R(x,a) = 0 \quad \text{and} \quad \sum_{y \in \mathcal{K}} p_{x\,y}(a) = 1, \quad x \in \mathcal{K},\ a \in A(x). \tag{3.2}$$

(ii) The models M and $\widetilde{M}$ are closely related; in fact, starting at the same state $x \in S \setminus \mathcal{K}$, the state-action processes $\{(X_t, A_t)\}$ and $\{(\widetilde{X}_t, \widetilde{A}_t)\}$ corresponding to the two models evolve in the same way, and the associated reward streams coincide, *while the systems stay in* $S \setminus \mathcal{K}$, but a transition into $\mathcal{K}$ in the model M corresponds to a transition into state Δ for the model $\widetilde{M}$, where the system remains forever earning a null reward. Thus, it can be said that $\widetilde{M}$ is obtained from M by "collapsing" the kernel $\mathcal{K}$ into a single absorbing state.

THEOREM 3.2. *Let* $\lambda \neq 0$ *and suppose that Assumptions* 2.1 *and* 2.2 *hold.*

(i) *For each* $x \in S \setminus \mathcal{K}$, $V_\lambda^*(x) = \widetilde{V}_\lambda^*(x)$.

(ii) *Let* f *and* $\widetilde{f}$ *be stationary policies for the models* M *and* $\widetilde{M}$, *respectively, which satisfy*

$$f(x) = \widetilde{f}(x), \qquad x \in S \setminus \mathcal{K}.$$

In this case, $V_\lambda(f, x) = \widetilde{V}_\lambda(\widetilde{f}, x)$ *for every* $x \in S \setminus \mathcal{K}$.

(iii) *The model* $\widetilde{M}$ *is unichain if the following condition holds* (*see* (2.9)):

$$P_f[T_{\mathcal{K}} < \infty \,|\, X_0 = x] = 1, \qquad f \in \mathbb{F},\ x \in S. \tag{3.3}$$

(iv) *If, in addition to Assumptions* 2.1 *and* 2.2, (3.3) *holds, then there exists a* λ*-optimal stationary policy for the model* M. *Moreover, if* $f \in \mathbb{F}$ *is obtained by maximizing the right-hand side of the* λ*-OE, i.e.,*

$$U_\lambda(V_\lambda^*(x)) = e^{\lambda R(x, f(x))} \sum_y p_{x\,y}(f(x)) U_\lambda(V_\lambda^*(y)), \qquad x \in S,$$

then f *is* λ*-optimal for the model* M.

This result follows, essentially, from the relation between the models M and $\widetilde{M}$ described in Remark 3.1. Since Theorem 3.2 plays a central role in this paper, a detailed proof will be provided.

Proof. (i) Define the functions $\widetilde{W} : \widetilde{S} \to [0, \infty)$ and $W : S \to [0, \infty)$ as follows:

$$\begin{aligned} &\widetilde{W}(x) = V_\lambda^*(x), \quad x \in S \setminus \mathcal{K}, \qquad \widetilde{W}(\Delta) = 0, \\ &W(x) = \widetilde{V}_\lambda^*(x), \quad x \in S \setminus \mathcal{K}, \qquad W(x) = 0, \quad x \in \mathcal{K}. \end{aligned} \tag{3.4}$$

From the optimality equation for the model M it follows that for each $x \in S$ and $a \in A(x)$,

$$\begin{aligned} U_\lambda(V_\lambda^*(x)) &\geq e^{\lambda R(x,a)} \sum_y p_{x\,y}(a) U_\lambda(V_\lambda^*(y)) \\ &= e^{\lambda R(x,a)} \Big[\sum_{y \in S \setminus \mathcal{K}} p_{x\,y}(a) U_\lambda(V_\lambda^*(y)) + \Big(\sum_{y \in \mathcal{K}} p_{x\,y}(a) \Big) U_\lambda(0) \Big]; \end{aligned}$$

for the equality, recall that $V_\lambda^*(x) = 0$ when $x \in \mathcal{K}$. Combining this expression with the definition of the components of the model $\widetilde{M}$ and the specification of $\widetilde{W}$ in (3.4), we deduce, for every $x \in S \setminus \mathcal{K}$ and $a \in \widetilde{A}(x)$,

$$U_\lambda(\widetilde{W}(x)) \geq e^{\lambda \widetilde{R}(x,a)} \sum_{y \in \widetilde{S}} \widetilde{p}_{x\,y}(a) U_\lambda(\widetilde{W}(y)).$$

Moreover, since the single action at state Δ, namely a_Δ, satisfies $\widetilde{R}(\Delta, a_\Delta) = 0$ and $\widetilde{p}_{\Delta\,\Delta}(a_\Delta) = 1$, the above inequality also holds for $x = \Delta$ and $a = a_\Delta$. Then, since $\widetilde{M}$ clearly satisfies Assumption 2.1, Lemma 2.2 applied to $\widetilde{M}$

yields that

$$\widetilde{W}(\cdot) \geq \widetilde{V}_\lambda^*(\cdot); \tag{3.5}$$

in particular, $\widetilde{V}_\lambda^*(\cdot)$ is a finite function, and the optimality equation for the model $\widetilde{M}$ yields that for every $x \in \widetilde{S}$ and $a \in \widetilde{A}(x)$,

$$\begin{aligned} U_\lambda(\widetilde{V}_\lambda^*(x)) &\geq e^{\lambda \widetilde{R}(x,a)} \sum_{y \in \widetilde{S}} \widetilde{p}_{x\,y}(a) U_\lambda(\widetilde{V}_\lambda^*(y)) \\ &= e^{\lambda \widetilde{R}(x,a)} \Big[\sum_{y \in S \setminus \mathcal{K}} \widetilde{p}_{x\,y}(a) U_\lambda(\widetilde{V}_\lambda^*(y)) + \widetilde{p}_{x\,\Delta}(a) U_\lambda(0) \Big]; \end{aligned}$$

where the equality used the fact that $\widetilde{V}_\lambda^*(\Delta) = 0$; recall that Δ is absorbing and that $R(\Delta, a_\Delta) = 0$. By Definition 3.1 and the specification of $W(\cdot)$ in (3.4), this yields that, for every $x \in S \setminus \mathcal{K}$ and $a \in A(x)$,

$$U_\lambda(W(x)) \geq e^{\lambda R(x,a)} \Big[\sum_{y \in S \setminus \mathcal{K}} p_{x\,y}(a) U_\lambda(W(y)) + \sum_{y \in \mathcal{K}} p_{x\,y}(a) U_\lambda(W(y)) \Big],$$

whereas, using (3.2), it is not difficult to verify that this inequality also holds for $x \in \mathcal{K}$ and $a \in A(x)$. Thus, Lemma 2.2 applied to the model M yields that $W(\cdot) \geq V_\lambda^*(\cdot)$, and the conclusion follows by combining this inequality with (3.4) and (3.5).

(ii) Let f and $\widetilde{f}$ be stationary policies for the models M and $\widetilde{M}$, respectively, and assume that $f(x) = \widetilde{f}(x)$ for $x \in S \setminus \mathcal{K}$. Consider the model $M_f = (S, A, \{A_f(x)\}, R, P)$ obtained from M by setting $A_f(x) = \{f(x)\}$ for every $x \in S$, so that $f(x)$ is the only available action at x for the model M_f. Similarly, let $\widetilde{M}_{\widetilde{f}}$ be the model obtained from $\widetilde{M}$ by restricting the admissible actions at each $x \in \widetilde{S}$ to the singleton $\{\widetilde{f}(x)\}$, i.e., $\widetilde{M}_{\widetilde{f}} = (\widetilde{S}, A, \{\widetilde{A}_{\widetilde{f}}(x)\}, \widetilde{R}, \widetilde{P})$, where $\widetilde{A}_{\widetilde{f}}(x) = \{\widetilde{f}(x)\}$ for each $x \in \widetilde{S}$. In this case, f and $\widetilde{f}$ are the *single* stationary policies for the models M_f and $\widetilde{M}_f$, so that the corresponding optimal value functions are $V_\lambda(f, \cdot)$ and $\widetilde{V}_\lambda(\widetilde{f}, \cdot)$, respectively. By observing that $V_\lambda(f, x) \leq V_\lambda^*(x) = 0$ for every $x \in \mathcal{K}$, the equality of $V_\lambda(f, \cdot)$ and $\widetilde{V}_\lambda(\widetilde{f}, \cdot)$ can be established along the same arguments used in the proof of part (i).

(iii) Let $\widetilde{f}$ be an arbitrary stationary policy for the model $\widetilde{M}$, and define $f \in \mathbb{F}$ by $f(x) = \widetilde{f}(x)$ if $x \in S \setminus \mathcal{K}$, and $f(x) = g(x)$ for $x \in \mathcal{K}$, where $g \in \mathbb{F}$ is arbitrary but fixed. It will be shown, by induction, that for every $n \in \mathbb{N}$,

$$P_{\widetilde{f}}[T_\Delta > n \,|\, \widetilde{X}_0 = x] = P_f[T_{\mathcal{K}} > n \,|\, X_0 = x], \quad x \in S \setminus \mathcal{K}. \tag{3.6}$$

For $n = 0$ both sides of this equality reduce to one; see (2.9). Assuming that (3.6) holds for $n = k \in \mathbb{N}$, observe that, by the Markov property, the

definition of the stopping times T_Δ and $T_{\mathcal{K}}$ yields

$$\begin{aligned} P_{\widetilde{f}}[T_\Delta > k+1 \,|\, \widetilde{X}_0 = x] &= \sum_{y \in S\setminus\mathcal{K}} \widetilde{p}_{x\,y}(\widetilde{f}(x)) P_{\widetilde{f}}[T_\Delta > k \,|\, \widetilde{X}_0 = y] \\ &= \sum_{y \in S\setminus\mathcal{K}} p_{x\,y}(f(x)) P_f[T_{\mathcal{K}} > k \,|\, X_0 = y] \\ &= P_f[T_{\mathcal{K}} > k+1 \,|\, X_0 = x] \end{aligned}$$

where the second equality used the induction hypothesis, as well as the specifications of policy f and the transition law $[\widetilde{p}_{x\,y}(\cdot)]$ in Defnition 3.1. Thus, (3.6) holds for every $n \in \mathbb{N}$, and taking limit as n goes to ∞, we deduce that $P_{\widetilde{f}}[T_\Delta = \infty \,|\, \widetilde{X}_0 = x] = P_f[T_{\mathcal{K}} = \infty \,|\, X_0 = x]$ for every $x \in S\setminus\mathcal{K}$. Under (3.3), it follows that $P_{\widetilde{f}}[T_\Delta = \infty \,|\, \widetilde{X}_0 = x] = 0$, so that

$$P_{\widetilde{f}}[T_\Delta < \infty \,|\, \widetilde{X}_0 = x] = 1, \qquad x \in S \setminus \mathcal{K}.$$

This equality shows that, under the action of policy $\widetilde{f}$, $\Delta \in \widetilde{S}$ is accessible from every state in $S \setminus \mathcal{K}$; since the set $\{\Delta\}$ is clearly minimal $\widetilde{f}$-closed, it follows that $\mathcal{E}(\widetilde{f}) = 1$ and then $\widetilde{M}$ is unichain.

(iv) Under (3.3), $\widetilde{M}$ is a unichain MDP (by part (iii)) which satisfies Assumptions 2.1 and 2.2. Therefore, by Lemma 2.1, there exists a λ-optimal stationary policy $\widetilde{f}$ for $\widetilde{M}$. If $f \in \mathbb{F}$ is such that $f(x) = \widetilde{f}(x)$ for $x \in S \setminus \mathcal{K}$, then parts (i) and (ii) yield that $V_\lambda(f,x) = \widetilde{V}_\lambda(\widetilde{f},x) = \widetilde{V}_\lambda^*(x) = V_\lambda^*(x)$ for every $x \in S \setminus \mathcal{K}$; since $0 \le V_\lambda(f,x) \le V_\lambda^*(x) = 0$ for $x \in \mathcal{K}$, it follows that f is λ-optimal for model M. To conclude, suppose that

$$U_\lambda(V_\lambda^*(x)) = e^{\lambda R(x,f(x))} \sum_{y \in S} p_{x\,y}(f(x)) U_\lambda(V_\lambda^*(y)), \qquad x \in S,$$

set $\widetilde{f}(x) = f(x)$ for $x \in S \setminus \mathcal{K}$, and $\widetilde{f}(\Delta) = a_\Delta$. In this case, straightforward calculations using Definition 3.1 show that

$$U_\lambda(\widetilde{V}_\lambda^*(x)) = e^{\lambda \widetilde{R}(x,\widetilde{f}(x))} \sum_{y \in \widetilde{S}} p_{x\,y}(\widetilde{f}(x)) U_\lambda(\widetilde{V}_\lambda^*(y)), \qquad x \in \widetilde{S};$$

since $\widetilde{M}$ is unichain, $\widetilde{f}$ is λ-optimal, by Lemma 2.1, and then, since $f(x) = \widetilde{f}(x)$ for $x \in S \setminus \mathcal{K}$, the above argument yields that f is λ-optimal for the original model M. ∎

According to Theorem 3.2, a λ-optimal stationary policy exists for a model M satisfying Assumptions 2.1 and 2.2 whenever condition (3.3) occurs; unfortunately, this latter requirement does not need to hold, even under the additional Assumption 3.1. However, the strategy to establish Theorem 3.1 will be based on Theorem 3.2, and can be described as follows: It

will be shown that there exist closed sets $\widehat{A}(x) \subset A(x)$ such that the new MDP $\widehat{M} = (S, A, \{\widehat{A}(x)\}, R, P)$ satisfies the following assertions (a) and (b):

(a) the optimal value functions of $\widehat{M}$ and M coincide, and

(b) condition (3.3) is satisfied for the model $\widehat{M}$.

In this case, there exists a λ-optimal stationary policy for the model $\widehat{M}$, which in turn is also optimal for the original MDP. Moreover, every stationary policy obtained by maximizing the right-hand side of the λ-OE for $\widehat{M}$ is λ-optimal for the model M.

4. Preliminaries. This section contains the technical tools that will be used to prove Theorem 3.1 as outlined above. The argument has been split into four simple parts presented below as Lemmas 4.1–4.4.

LEMMA 4.1. *Let $f \in \mathbb{F}$ and $G \subset S$ be arbitrary, and suppose that*

$$G \cap C^* \neq \emptyset \quad \text{for every minimal } f\text{-closed set } C^*. \tag{4.1}$$

In this case, for every $x \in S$,

$$P_f[T_G < \infty \,|\, X_0 = x] = 1. \tag{4.2}$$

Proof. Let C^* be a given minimal f-closed set. Using (4.1), pick $y \in G \cap C^*$ and observe that the definition of T_G and Remark 2.1(i) together yield that

$$P[T_G < \infty \,|\, X_0 = x] \geq P_f[X_n = y \text{ for some } n \in \mathbb{N} \setminus \{0\} \,|\, X_0 = x] = 1, \quad x \in C^*,$$

so that (4.2) is certainly valid when $x \in C^*$. Since the class $\mathcal{R}(f)$ of recurrent states with respect to the Markov chain induced by f is the union of minimal f-closed sets (see Remark 2.1(i) again), it follows that

$$P_f[T_G < \infty \,|\, X_0 = x] = 1, \quad x \in \mathcal{R}(f). \tag{4.3}$$

On the other hand, it is well known that, since the state space is finite,

$$P[T_{\mathcal{R}(f)} < \infty \,|\, X_0 = x] = 1, \quad x \in S;$$

see, for instance, Billingsley (1995), or Loève (1977). Thus, via the Markov property, (4.2) follows by combining this latter equality and (4.3). ∎

LEMMA 4.2. *Suppose that $f \in \mathbb{F}$ is such that $V_\lambda(f, x) < \infty$ for every $x \in S$. In this case*

$$V_\lambda(f, x) = 0, \quad x \in \mathcal{R}(f).$$

Proof. Let C^* be an arbitrary minimal f-closed set. As already noted in Remark 2.1, each pair of states in C^* communicate under the action of

f, and each member of C^* is positive recurrent under the action of f. Then (Chapter 3 of Loève (1977), Section 8 of Billingsley (1995)) we have,

$$P_f[X_n = y \text{ for an infinite number of integers } n \,|\, X_0 = x] = 1, \quad x, y \in C^*,$$

and since the reward function is nonnegative and the utility function is strictly increasing, this yields that for each pair $x, y \in C^*$ and $m \in \mathbb{N}$,

$$\begin{aligned} U_\lambda(V_\lambda(f,x)) &= E_f\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(X_t, A_t)\Big) \,\Big|\, X_0 = x\Big] \\ &\geq E_f\Big[U_\lambda\Big(\sum_{t=0}^{\infty} R(y, f(y)) I[X_t = y]\Big) \,\Big|\, X_0 = x\Big] \\ &\geq E_f[U_\lambda(mR(y, f(y))) \,|\, X_0 = x] \\ &= U_\lambda(mR(y, f(y))), \end{aligned}$$

where $I[X_t = y]$ denotes the indicator function of the event $[X_t = y]$. Thus, $V_\lambda(f,x)) \geq mR(y, f(y))$; since $m \in \mathbb{N}$ is arbitrary and $R(\cdot,\cdot) \geq 0$, this yields that $R(y, f(y))$ for each $y \in C^*$. Therefore, since C^* is an f-closed set it follows that $P_f[R(X_t, A_t) = 0 \,|\, X_0 = x] = 1$, for every $t \in \mathbb{N}$ and $x \in C^*$, so that

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

and then $V_\lambda(f,x) = 0$ for each $x \in C^*$. The conclusion follows from recalling that $\mathcal{R}(f)$ is the union of all the minimal f-closed sets. ∎

The next step to prove Theorem 3.1 is the following result, which ensures the existence of ε-optimal policies in the relative sense.

LEMMA 4.3. *Fix $\varepsilon \in (0,1)$, and suppose that Assumptions* 4.1 *and* 4.2 *hold. In this case, there exists a policy $f \in \mathbb{F}$ such that*

$$V_\lambda(f,x) \geq (1-\varepsilon)V_\lambda^*(x), \quad x \in S.$$

This lemma was recently established in Cavazos-Cadena and Montes-de-Oca (2000c), where the argument relies on a strong form of the λ-OE in (2.7); another proof, using *a discounted* dynamic programming operator, can be found in Cavazos-Cadena and Montes-de-Oca (2000b). The remainder of the section involves a construction based on Lemma 4.3 which, essentially, looks for a policy $f \in \mathbb{F}$ which is a "good candidate" for λ-optimality. First, for each $n \in \mathbb{N}$, Lemma 4.3 yields the existence of a policy $f_n \in \mathbb{F}$ satisfying

$$V_\lambda(f_n, \cdot) \geq \Big(1 - \frac{1}{n+2}\Big) V_\lambda^*(\cdot); \tag{4.4}$$

since the space $\mathbb{F}$ is compact metric, without loss of generality it can be assumed, by taking a subsequence if necessary, that

$$\lim_{n\to\infty} f_n = f \in \mathbb{F}. \tag{4.5}$$

Moreover, since $\mathcal{E}(f_n)$ is an integer number less than or equal to the number of states, an additional subsequence can be taken so that, for a fixed integer k,

$$\mathcal{E}(f_n) = k, \quad n \in \mathbb{N}. \tag{4.6}$$

LEMMA 4.4. *Suppose that Assumptions* 2.1, 2.2 *and* 3.1 *hold. With the notation in* (4.4)–(4.6), *the following assertions* (i)–(ii) *are true*:

(i) $\mathcal{E}(f) = k$, *and*
(ii) $\mathcal{R}(f) \subset \mathcal{K}$;

Proof. Part (i) follows from (4.6) and the continuity of $\mathcal{E}(\cdot)$. To establish part (ii), let $C_i^*(f_n)$, $i = 1, \ldots, k$, be the k disjoint minimal f_n-closed sets, and observe that Lemma 4.2 and (4.4) together imply that

$$\bigcup_{i=1}^{k} C_i^*(f_n) = \mathcal{R}(f_n) \subset \mathcal{K}, \quad n \in \mathbb{N}. \tag{4.7}$$

Consider the Markov chain induced by f_n, and for each $n \in \mathbb{N}$ and $i = 1, \ldots, k$, let $\mu_{i,n}$ be the unique invariant distribution supported in $C_i^*(f_n)$, so that the following assertions (a)–(c) are satisfied for each $n \in \mathbb{N}$:

(a) $\mu_{i,n}(y) = \sum_x \mu_{i,n}(x) p_{x\,y}(f_n(x))$, $y \in S$, $i = 1, \ldots, k$.
(b) $\sum_{x\in\mathcal{K}} \mu_{i,n}(x) = 1$, $i = 1, \ldots, k$

(notice that this equality follows from the fact that $\operatorname{supp}(\mu_{i,n}) = C_i^*(f_n) \subset \mathcal{K}$; see (4.7)).

(c) $\sum_{x\in\mathcal{K}} |\mu_{i,n}(x) - \mu_{j,n}(x)| = 2$ if $i \neq j$

(this assertion is equivalent to the statement that the supports of $\mu_{i,n}$ and $\mu_{j,n}$, namely $C_i^*(f_n)$ and $C_j^*(f_n)$, are disjoint when $i \neq j$).

Since the state space is finite, for each $i = 1, \ldots, k$ the sequence $\{\mu_{i,n}\}_{n\in\mathbb{N}}$ is tight. Therefore, by taking a subsequence if necessary, it can be assumed that the following convergences hold, where $\mu_i(\cdot)$, $i = 1, \ldots, k$, are probability distributions on S:

$$\lim_{n\to\infty} \mu_{i,\,n}(\cdot) = \mu_i(\cdot), \quad i = 1, \ldots, k.$$

When we combine these convergences with properties (a)–(c) above, the following assertions follow via (4.5) and the continuity of the transition law:

(a′) For each $i = 1, \ldots, k$, $\mu_i(y) = \sum_x \mu_i(x) p_{x\,y}(f(x))$, $y \in S$.
(b′) $\sum_{x\in\mathcal{K}} \mu_i(x) = 1$;
(c′) $\sum_{x\in\mathcal{K}} |\mu_i(x) - \mu_j(x)| = 2$ if $i \neq j$.

Thus, each μ_i is an invariant distribution of the Markov chain induced by f, and via (b′), it follows that

$$\operatorname{supp}(\mu_i) \subset \mathcal{K}, \quad i = 1, \ldots, k,$$

whereas (c′) immediately yields that

$$\operatorname{supp}(\mu_i) \cap \operatorname{supp}(\mu_j) = \emptyset \quad \text{if } i \neq j.$$

As observed in Remark 2.1, for each $i = 1, \ldots, k$, there exists a minimal f-closed set C_i^* contained in $\operatorname{supp}(\mu_i)$, so that the last two displayed relations together imply that

$$C_i^* \subset \mathcal{K}, \quad i = 1, \ldots, k, \quad \text{and} \quad C_i^* \cap C_j^* = \emptyset \quad \text{when } i \neq j.$$

Since $\mathcal{E}(f) = k$, these sets C_i^* are *all* the k different minimal f-closed sets, and the inclusion in the last displayed relation implies that $\mathcal{R}(f) = C_1^* \cup C_2^* \cup \ldots \cup C_k^* \subset \mathcal{K}$, completing the proof of part (ii). ∎

5. Proof of the main result. In this section a proof of Theorem 3.1 will be provided. For the sake of clarity, the essential part of the argument is stated separately in the following lemma.

LEMMA 5.1. *Suppose that Assumption* 3.1 *holds, and let $f \in \mathbb{F}$ be such that $\mathcal{R}(f) \subset G$, where $G \subset S$ is fixed. In this case, for every $x \in S$ there exists a closed neighborhood $B(f(x)) \subset A$ of $f(x)$ such that*

$$f' \in \prod_{x \in S} (B(f(x)) \cap A(x)) \Rightarrow C^* \cap G \neq \emptyset \text{ for each minimal } f'\text{-closed set } C^*. \tag{5.1}$$

Proof. The argument is by contradiction. First, notice that, by Remark 2.1(i), the inclusion $\mathcal{R}(f) \subset G$ is equivalent to the assertion that *all the minimal f-closed sets are contained in G*, and let $B_n(f(x))$ be the closed ball in A of radius $1/(n+1)$ with center $f(x)$. Suppose that for each $n \in \mathbb{N}$, it is possible to find a stationary policy f_n such that

$$f_n \in \prod_{x \in S} (B_n(f(x)) \cap A(x)) \quad \text{and} \quad C_n^* \subset S \setminus G, \tag{5.2}$$

where C_n^* is a minimal f_n-closed set. Let μ_n be the invariant probability distribution of the Markov chain induced by f_n associated with C^*, i.e., $\mu_n(\cdot)$ satisfies

$$\mu_n(y) = \sum_x \mu_n(x) p_{x\,y}(f_n(x)), \quad y \in S, \tag{5.3}$$

and $\operatorname{supp}(\mu_n) = C_n^* \subset S \setminus G$, so that

$$\sum_{x \in S \setminus G} \mu_n(x) = 1. \tag{5.4}$$

Since the state space is finite, the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ is tight and, for some subsequence $\{\mu_{n_k}\}$, the following convergence holds for a probability distribution μ on S:

$$\lim_{k\to\infty} \mu_{n_k}(x) = \mu(x), \qquad x \in S. \tag{5.5}$$

On the other hand, since $B_n(f(x))$ is a closed ball with radius $1/(n+1)$, (5.2) implies that $\lim_{n\to\infty} f_n = f$, and the continuity of the transition law and (5.3)–(5.5) together yield that

$$\mu(y) = \sum_x \mu(x) p_{x\,y}(f(x)), \qquad y \in S,$$

and

$$\sum_{x\in S\setminus G} \mu(x) = 1.$$

Thus, μ is an invariant distribution of the Markov chain induced by f, whose support is contained in $S \setminus G$. Next, recall that $\mathrm{supp}(\mu)$ contains a minimal f-closed set C^* (see Remark 2.1(iii)), and in this case $C^* \subset \mathrm{supp}(\mu) \subset S\setminus G$; this is a contradiction since, as already noted, all the minimal f-closed sets are contained in G. Thus, (5.1) holds when, for each $x \in S$, $B(f(x)) = B_n(f(x))$ with n large enough. ∎

Proof of Theorem 3.1. Let the stationary policy f be as in Lemma 4.4, so that f is the limit of a sequence $\{f_n\}$ satisfying (4.4). By Lemma 4.4(ii), $\mathcal{R}(f) \subset \mathcal{K}$. Using Lemma 5.1 with this policy f and the kernel $\mathcal{K}$ instead of G, for each $x \in S$ select a closed neighborhood $B(f(x)) \subset A$ of $f(x)$ such that (5.1) holds, and define the new MDP $\widehat{M} = (S, A, \{\widehat{A}(x)\}, R, P)$ by setting $\widehat{A}(x) = B(f(x)) \cap A(x)$, $x \in S$; notice that each set $\widehat{A}(x)$ is a compact subset of $A(x)$ and that the class $\widehat{\mathcal{P}}$ of admissible policies for the model $\widehat{M}$ is contained in $\mathcal{P}$, so that

$$\widehat{V}^*_\lambda(x) = \sup_{\pi\in\widehat{\mathcal{P}}} V_\lambda(\pi, x) \le \sup_{\pi\in\mathcal{P}} V_\lambda(\pi, x) = V^*_\lambda(x), \qquad x \in S. \tag{5.6}$$

On the other hand, since each set $B(f(x))$ is a (closed) *neighborhood* of $f(x)$, the convergence in (4.5) implies that f_n is an admissible stationary policy for $\widehat{M}$ whenever n is large enough, so that (4.4) implies that $\widehat{V}^*_\lambda(\cdot) \ge V^*_\lambda(\cdot)$, and combining this with (5.6) we deduce that

$$\widehat{V}^*_\lambda(\cdot) = V^*_\lambda(\cdot). \tag{5.7}$$

To conclude observe that (5.1) implies that every $\widehat{f} \in \prod_{x\in S} \widehat{A}(x)$ satisfies the condition (4.1) with $G = \mathcal{K}$, so Lemma 4.1 implies that

$$P_{\widehat{f}}[T_{\mathcal{K}} < \infty \,|\, X_0 = x] = 1, \qquad x \in S,\ \widehat{f} \in \prod_{x\in S} \widehat{A}(x).$$

Thus, by Theorem 3.2(iv) applied to the model $\widehat{M}$, there exists a policy $\widehat{f} \in \prod_{x\in S} \widehat{A}(x)$ such that $\widehat{V}_\lambda(\widehat{f},\cdot) = \widehat{V}^*_\lambda(\cdot)$, and then $\widehat{f}$ is also optimal for the model M, by (5.7). Moreover, every stationary policy obtained by maximizing the right-hand side of the optimality equation corresponding to $\widehat{M}$ is optimal for both models $\widehat{M}$ and M. ∎

6. Conclusion. This work considered *multichain* MDPs with finite state space and nonnegative rewards. Under the structural stability condition in Assumption 3.1 and the continuity-compactness requirements in Assumption 2.1, it was shown that an optimal stationary policy exists whenever the optimal value function is finite. In contrast with the usual approach in the risk-neutral case, which is based on the discounted criterion (Schäl (1984), (1986)), in the present framework Theorem 3.1 was derived via a reduction to a unichain MDP, for which the existence of an optimal stationary policy had already been established. Finally, trying to extend Theorem 3.1 to more general contexts, for instance, MDPs with more general state space or rewards with unrestricted sign, seems to be an interesting problem; research in this direction is currently in progress.

References

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

P. Billingsley (1995), *Probability and Measure*, Wiley, New York.

R. Cavazos-Cadena, E. Feinberg and R. Montes-de-Oca (2000), *A note on the existence of optimal policies in total reward dynamic programs with compact action sets*, Math. Oper. Res., in press.

R. Cavazos-Cadena and R. Montes-de-Oca (2000a), *Optimal stationary policies in risk-sensitive dynamic programs with finite state space and nonnegative rewards*, Appl. Math. (Warsaw) 27, 167–185.

R. Cavazos-Cadena and R. Montes-de-Oca (2000b), *Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces*, Math. Methods Oper. Res. 52, 133–167.

R. Cavazos-Cadena and R. Montes-de-Oca (2000c), *Optimal and nearly optimal policies in Markov decision chains with nonnegative rewards and risk-sensitive expected total-reward criterion*, in: Markov Processes and Controlled Markov Chains, J. Filar, H. Zhenting and A. Chen (eds.), to appear.

J. Dugundji (1966), *Topology*, Allyn and Bacon, Boston.

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

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 Oper. Res. 33, Springer, New York.

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, no. 1, 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.

M. Schäl (1984), *Markovian decision models with bounded finite-state rewards*, Operations Research Proceedings 1983, Springer, Berlin, 470–473.

M. Schäl (1986), *Markov and semi-Markov decision models and optimal stopping*, in: Semi-Markov Models, J. Janssen (ed.), Plenum Press, New York, 39–62.

P. J. Schweitzer (1968), *Perturbation theory and finite Markov chains*, J. Appl. Probab. 5, 401–413.

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

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 10.12. 1999;
revised version on 20.4.2000 (1517)