Appl Math Optim 40:273–285 (1999)

Applied Mathematics & Optimization

Existence of Risk-Sensitive Optimal Stationary Policies for Controlled Markov Processes*

D. Hernández-Hernández[1] and S. I. Marcus[1,2]

[1]Institute for Systems Research, University of Maryland,
College Park, MD 20742, USA
marcus@isr.umd.edu

[2]Electrical Engineering Department, University of Maryland,
College Park, MD 20742, USA

Communicated by A. Bensoussan

Abstract. In this paper we are concerned with the existence of optimal stationary policies for infinite-horizon risk-sensitive Markov control processes with denumerable state space, unbounded cost function, and long-run average cost. Introducing a discounted cost dynamic game, we prove that its value function satisfies an Isaacs equation, and its relationship with the risk-sensitive control problem is studied. Using the vanishing discount approach, we prove that the risk-sensitive dynamic programming inequality holds, and derive an optimal stationary policy.

Key Words. Risk-sensitive stochastic control, Dynamic games, Isaacs equation, Optimal stationary policies.

AMS Classification. 90C40 (93E20).

1. Introduction

In this paper we are concerned with the existence of optimal stationary policies for infinite-horizon risk-sensitive stochastic control problems with denumerable state space, discrete time parameter, unbounded cost function, and long-run average cost. For the risk-neutral stochastic control problem, the same kind of problem has been addressed, see,

* This research was supported in part by the National Science Foundation under Grant EEC 9402384. The first author is on leave from the Department of Mathematics, CINVESTAV-IPN, Mexico.

e.g., [C], [CS], [S1], [S2], [HL2], and [HL3], exploiting the vanishing discount approach, in which the value function of the average cost control problem is approximated by the value function of a sequence of discounted problems. However, for the risk-sensitive control problem there does not seem to be a sequence of discounted control problems with which we can approximate the value function of the average cost problem. Therefore, we introduce a dynamic game, and consider both the discounted and the average cost criteria. Establishing some relationships (see Theorem 3.1) between the value function of the average cost dynamic game and the value function of the risk-sensitive control problem, it is possible to approximate the value function of the risk-sensitive control problem through the value function of a discounted cost dynamic game, which satisfies an Isaacs equation. Then, using well-known techniques of the vanishing discount approach, we prove the existence of a solution to the risk-sensitive dynamic programming inequality (DPI), and derive an optimal stationary policy. In [HM] it was proved that there exists a bounded solution to the risk-sensitive dynamic programming equation (DPE), under conditions that force the controlled process to have very strong recurrence properties for all stationary policies. In this paper we introduce weaker assumptions, and prove the existence of a solution to the DPI.

The use of game theory to solve this problem is not surprising, and it has been explored extensively in the study of risk-sensitive control problems [BJ], [FH], [FM1], [FM2], [DMR], [W]. See also [FGM], where risk-sensitive control problems for hidden Markov models were treated. A key tool for establishing the relationships between dynamic games and the risk-sensitive control problem is a variational lemma, that expresses the duality relationship between the relative entropy function and the logarithmic moment-generating function. Recently, Dupuis and Ellis [DE] found interesting applications of this lemma in their study of representation formulas and weak convergence methods.

The paper is organized as follows. Section 2 describes the control model we deal with. In Section 3 we introduce some preliminary results, and finally Section 4 contains the main result.

2. Preliminaries

The Control Model. Let (S, A, π, c) be a Markov control model [ABF$^+$], [HL1] satisfying the following. The set $S = \{0, 1, \ldots\}$ is the state space, endowed with the discrete topology, while A is a Borel space, called the action or control space. For every $x \in S$, there is a nonempty set $A(x) \subset A$, which represents the set of admissible actions when the system is in state x. The set of admissible pairs is $\mathbf{K} := \{(x, a) : x \in S, a \in A(x)\}$, and is assumed to be a Borel subspace of $S \times A$. The transition law π is a stochastic kernel on S given $\mathbf{K}$. Finally, c: $\mathbf{K} \to \mathbb{R}$ is a lower semicontinuous (l.s.c.) function, nonnegative, which represents the one stage cost.

Assumption A.1.

(i) For each $x, y \in S$, the mapping $a \to \pi(y|x, a)$, with $a \in A(x)$, is l.s.c.
(ii) For each $x \in S$, $A(x)$ is a compact subset of A.

Define $H_0 = S$, and $S_t = \mathbf{K} \times H_{t-1}$ if $t = 1, 2, \ldots$. A control policy, or strategy, is a sequence $\vec{\delta} = \{\delta_t\}$ of stochastic kernels on A given H_t that satisfy the constraint

$$\delta_t(A(x_t)|h_t) = 1, \qquad \forall h_t \in H_t, \quad t \geq 0.$$

The set of policies is denoted by Δ. A policy $\vec{\delta} \in \Delta$ is called a Markov policy if there exists a sequence of functions $\{\pi_t\}$, with $\pi_t\colon\ S \to P(A)$, where $P(A)$ is the set of probability measures on A, such that $\pi_t(x)(A(x)) = 1$. We denote by Δ_M the set of Markov policies, and throughout we restrict ourselves, without loss of generality, to this set of control policies. We denote by $\mathbf{F}$ the set of functions $f\colon\ S \to A$ such that $f(x) \in A(x)$ for all $x \in S$. A policy $\vec{\delta} \in \Delta$ is stationary if there exists $f \in \mathbf{F}$ such that $\delta_t(f(x_t)|h_t) = 1$ for all $h_t \in H_t$, $t \geq 0$; this policy is also denoted by $f \in \mathbf{F}$.

If the initial state $x \in S$ and $\vec{\delta} \in \Delta_M$ are given, there exists a unique probability measure $P_x^{\vec{\delta}}$ on (Ω, ζ), the canonical measurable space that consists of the sample space $\Omega := (S \times A)^\infty$ and the corresponding product σ-algebra ζ. Further, a stochastic process $\{(x_t, a_t), t = 0, 1, \ldots\}$ is defined in a canonical way, where x_t and a_t denote the state and action at time t, respectively. The expectation operator with respect to $P_x^{\vec{\delta}}$ is denoted by $E_x^{\vec{\delta}}$.

Next we introduce the risk-sensitive cost criterion. For $x \in S, \vec{\delta} \in \Delta_M$, the cost functional to be minimized is defined by

$$J(x, \vec{\delta}) = \limsup_{T\to\infty} \gamma \frac{1}{T} \log E_x^{\vec{\delta}} \exp \left\{ \frac{1}{\gamma} \sum_{t=0}^{T-1} c(x_t, a_t) \right\},$$

where $\gamma > 0$ is the risk factor. Throughout, without loss of generality, we set $\gamma = 1$. Let

$$J(x) = \inf_{\Delta_M} J(x, \vec{\delta})$$

be the corresponding value function. Then the problem we are concerned with is to find a policy $f \in \mathbf{F}$ such that

$$J(x) = J(x, f^*).$$

Assumption A.2.

(a) There exists a stationary policy $\bar{f} \in \mathbf{F}$ such that

$$\rho := J(x, \bar{f})$$

is finite and independent of x.

(b)

$$\liminf_{x\to\infty} \min_{a\in A(x)} c(x, a) > \rho.$$

Remark 2.1. Assumption A.2 is a slight variation of that used in previous literature for the risk-neutral average cost criterion [C],[CS], [B]. However, the way we approach our problem is technically different, and depends heavily on the introduction of a dynamic game. This idea has been used in [HM], where dynamic programming techniques were used to prove the existence of optimal solutions to the risk-sensitive stochastic control problem with bounded cost function, and in [FH] for finite state problems.

In the remainder of this section we give a sufficient condition for Assumption A.2(a). See Theorem 2.1.10 of [DS]. Let $\bar{f} \in \mathbf{F}$, and let $\bar{x}_t$ be the Markov chain with transition kernel $\pi(y|x, \bar{f}(x))$.

Let $P(S)$ be the set of probability vectors on S, i.e.,

$$P(S) := \left\{ \mu = (\mu^0, \mu^1, \ldots) : \mu^i \geq 0, \sum_{i=0}^{\infty} \mu^i = 1 \right\},$$

endowed with the weak topology. We denote by Y^t the occupation measure of the Markov chain $\bar{x}_t$ with initial condition x, and assume that $\{Y^t\}$ satisfies the Large Deviation Principle in $P(S)$ with rate function independent of x. Further, let $\Phi : P(S) \to [0, \infty]$ be defined by

$$\Phi(\mu) = \sum_{x \in S} c(x, \bar{f}(x))\mu(x).$$

If Φ is finite, continuous, and satisfies, for each $x \in S$,

$$\lim_{C \to \infty} \limsup_{t \to \infty} \frac{1}{t} \log E_x^{\bar{f}} \{ 1_{\{\mu : \Phi(\mu) \geq C\}}(Y^t) \exp[t\Phi(Y^t)] \} = -\infty,$$

then, according to Theorem 2.1.10 of [DS], $\bar{f}$ satisfies Assumption A.2(a).

3. Stochastic Dynamic Games

We fix $\nu \in P(S)$. The relative entropy function $I(\cdot||\nu)$ is a map from $P(S)$ into the extended real numbers. It is defined by

$$I(\mu||\nu) := \begin{cases} \sum_{x \in S} \log(r(x))\mu(x) & \text{if} \quad \mu \ll \nu, \\ +\infty & \text{otherwise,} \end{cases}$$

where

$$r(x) = \begin{cases} \mu(x)/\nu(x) & \text{if} \quad \nu(x) \neq 0, \\ 1 & \text{otherwise.} \end{cases}$$

The stochastic dynamic game is defined as follows (see [FH] and [HM]). The set S is the state space, while A and $P(S)$ are the control sets for Players 1 and 2, respectively. The reward function is $(x, a, \mu) \to c(x, a) - I(\mu||\pi(\cdot|x, a))$, with $(x, a, \mu) \in \mathbf{K} \times P(S)$.

The evolution of the system is as follows. Let x_t be the state at time $t \in \{0, 1, \ldots\}$, and let a_t and μ_t be the actions chosen by Players 1 and 2, respectively. Then a reward

$c(x_t, a_t) - I(\mu_t||\pi(\cdot||x_t, a_t))$ is earned, and the system moves to the next state x_{t+1} according to the probability distribution μ_t.

For each $t \geq 0$, let $\mathbf{N}_t$ and $\mathbf{K}_t$ be the set of feasible histories up to time t for Players 1 and 2, respectively. That is, $\mathbf{N}_0 = S$ and $\mathbf{N}_t = (S \times P(S))^t \times S$, while $\mathbf{K}_0 = \mathbf{K}$ and $\mathbf{K}_t = \mathbf{K}^t \times \mathbf{K}$. We say that $\vec{f}$ is stationary if, for all $t \geq 0$, $f_t = f \in \mathbf{F}$ is independent of t. A randomized Markov strategy for Player 1 is a sequence $\vec{\delta} = \{\delta_t\}$ of functions from S to $P(A)$, such that $\delta_t\ (x)(A(x)) = 1$; with some abuse of notation, we denote this set of strategies as Δ_M. A nonrandomized Markov strategy for Player 1 is a sequence $\vec{f} = \{f_t\}$ of functions f_t from S to A, such that $f_t(x) \in A(x)$. A nonrandomized Markov strategy for Player 2 is a sequence $\vec{\xi} = \{\xi_t\}$ of stochastic kernel ξ_t on S given $\mathbf{K}$. Analogously, $\vec{\xi}$ is stationary if, for all $t \geq 0$, $\xi_t = \xi$: $\mathbf{K} \to P(S)$.

Let (Ω, ζ) be the canonical measurable space. Given the initial state $x \in S$, and strategies $\vec{\delta}, \vec{\xi}$, there exists a unique probability measure $P_x^{\vec{\delta},\vec{\xi}}$ and, again, a stochastic process $\{x_t, a_t, t \geq 0\}$ is defined on (Ω, ζ) in a canonical way, where x_t denotes the state at time t of the system, and a_t is the action for Player 1. The corresponding expectation operator is denoted by $E_x^{\vec{\delta},\vec{\xi}}$.

Given $x \in S, \vec{\delta}, \vec{\xi}$, define the cost functional

$$V_\beta(x, \vec{\delta}, \vec{\xi}) := E_x^{\vec{f},\vec{\xi}} \sum_{t=0}^{\infty} \beta^t [c(x_t, a_t) - I(\xi_t||\pi(\cdot|x_t, a_t))], \tag{3.1}$$

where $\beta \in (0, 1)$ is the discount factor. Note that, since c is (possibly) unbounded, $V_\beta(x, \vec{\delta}, \vec{\xi})$ might be undetermined. To avoid this, we restrict the set of admissible strategies for the second player in the following way. Given $\vec{\delta} \in \Delta_M$, we say that $\vec{\xi}$ is $\vec{\delta}$-admissible if $\vec{\xi}$ is a nonrandomized Markov strategy for Player 2, and, for each $x \in S$,

$$E_x^{\vec{\delta},\vec{\xi}} \sum_{t=0}^{T-1} I\ (\xi_t(\cdot|x_t, a_t)\|\pi(\cdot|x_t, a_t)) < \infty, \qquad T \geq 1,$$

and

$$\limsup_{T\to\infty} \frac{1}{T} E_x^{\vec{\delta},\vec{\xi}} \sum_{t=0}^{T-1} I\,(\xi_t(\cdot|x_t, a_t)\|\pi(\cdot|x_t, a_t)) < \infty.$$

We denote this set by $Q(\vec{\delta})$. Note that this set is not empty; $\xi = \pi \in Q(\vec{\delta})$. We define, analogously, the value function with average optimality criterion. Given $x \in S, \vec{\delta}, \vec{\xi}$, we define

$$\Lambda(x, \vec{\delta}, \vec{\xi}) := \limsup_{T\to\infty} \frac{1}{T} E_x^{\vec{\delta},\vec{\xi}} \sum_{t=0}^{T-1} [c(x_t, a_t) - I(\xi_t||\pi(\cdot|x_t, a_t))]. \tag{3.2}$$

Finally, we define the upper values of these games, respectively, by

$$V_\beta(x) := \inf_{\vec{\delta}} \sup_{\vec{\xi}\in Q(\vec{\delta})} V_\beta(x, \vec{\delta}, \vec{\xi})$$

and

$$\Lambda^*(x) := \inf_{\vec{\delta}} \sup_{\vec{\xi} \in Q(\vec{\delta})} \Lambda(x, \vec{\delta}, \vec{\xi}).$$

The following theorem is the basis for the existence of bounds which are used in the vanishing discount method.

Theorem 3.1. *Fix $T > 0$ and $\vec{\delta} \in \Delta_M$. For each $k = 0, \dots, T-1$ define*

$$\Lambda_{k,T-1}(x, \vec{\delta}) := \sup_{\vec{\xi} \in Q(\vec{\delta})} E_x^{\vec{\delta},\vec{\xi}} \left[\sum_{t=k}^{T-1} (c(x_t, a_t) - I(\xi_t || \pi(\cdot | x_t, a_t))) | x_k = x \right]$$

and

$$J_{k,T-1}(x, \vec{\delta}) = \log E_x^{\vec{\delta}} \exp \left[\sum_{t=k}^{T-1} c(x_t, a_t) | x_k = x \right].$$

Then

(a) *for all $x \in S$ and $k = 0, \dots, T-1$,*

$$\Lambda_{k,T-1}(x, \vec{\delta}) \le J_{k,T-1}(x, \vec{\delta}), \tag{3.3}$$

(b) $\limsup_{T\to\infty} (1/T) \Lambda_{0,T}(x, \vec{\delta}) \le J(x, \vec{\delta})$,
(c) $\Lambda^*(x) \le J(x)$.

Proof. We first prove (3.3) for $k = T-1$. Given $x \in S$, we assume that $J_{T-1,T-1}(x, \vec{\delta}) < \infty$, since otherwise (3.3) is obvious. Then

$$\begin{aligned} \Lambda_{T-1,T-1}(x, \vec{\delta}) &= \sup_{\vec{\xi} \in Q(\vec{\delta})} \int \left[c(x, a) - \int \log \left[\frac{d\xi_1(y|x, a)}{d\pi(y|x, a)} \right] \right. \\ &\qquad\qquad \left. \times\, \xi_1[dy|x, a] \delta_1(da|x) \right] \\ &\le \int c(x, a) \delta_{T-1}(da|x) \\ &\le J_{T-1,T-1}(x, \vec{\delta}). \end{aligned}$$

Now, we assume that (3.3) holds for $k = n+1, \dots, T-1$. Let $x \in S$ be such that $J_{n,T-1}(x, \vec{\delta}) < \infty$, and choose any $\vec{\xi} \in Q(\vec{\delta})$ such that

$$\Lambda_{n,T-1}(x, \vec{\delta}, \vec{\xi}) := E_x^{\vec{\delta},\vec{\xi}} \left[\sum_{t=n}^{T-1} [c(x_t, a_t) - I(\xi_t || \pi(\cdot | x_t, a_t))] | x_n = x \right]$$

is nonnegative. Then

$$
\begin{aligned}
\Lambda_{n,T-1}(x,\vec{\delta},\vec{\xi}) &= E_x^{\vec{\delta},\vec{\xi}}\Big[c(x_n,a_n) - I(\xi_n||\pi(\cdot|x_n,a_n)) \\
&\qquad + \int \Lambda_{n+1,T-1}(y,\vec{\delta},\vec{\xi})\xi_n(dy|x_n,a_n)|x_n = x\Big] \\
&\leq E^{\vec{\delta},\vec{\xi}}\Big[c(x_n,a_n) - I(\xi_n||\pi(\cdot|x_n,a_n)) \\
&\qquad + \int \Lambda_{n+1,T-1}(y,\vec{\delta})\xi_n(dy|x_n,a_n)|x_n = x\Big] \\
&\leq E^{\vec{\delta},\vec{\xi}}\Big[c(x_n,a_n) - I(\xi_n||\pi(\cdot|x_n,a_n)) \\
&\qquad + \int J_{n+1,T-1}(y,\vec{\delta})\xi_n(dy|x_n,a_n)|x_n = x\Big] \\
&= \int \Big[c(x,a) - I(\xi_n||\pi(\cdot|x,a)) \\
&\qquad + \int J_{n+1,T-1}(y,\vec{\delta})\xi_n(dy|x,a)\Big]\delta_n(da|x) \\
&\leq \int \Big[\log \int e^{c(x,a)+J_{n+1,T-1}(y,\vec{\delta})}\pi(dy|x,a)\Big]\delta_n(da|x) \\
&\leq J_{n,T-1}(x,\vec{\delta}),
\end{aligned}
$$

where the last inequality is due to Jensen's inequality. The proof of (b) follows immediately from (a). Now we prove (c). Let $\delta \in \Delta_M$, and choose $\vec{\xi} \in Q(\vec{\delta})$ such that $\Lambda(x,\vec{\delta},\vec{\xi}) \geq 0$. We prove first that

$$
\Lambda(x,\vec{\delta},\vec{\xi}) \leq J(x,\vec{\delta}). \tag{3.4}
$$

Assume that $J(x,\vec{\delta}) < \infty$, since otherwise there is nothing to prove. We first prove that $\Lambda(x,\vec{\delta},\vec{\xi}) < \infty$. Assume that $\Lambda(x,\vec{\delta},\vec{\xi}) = \infty$, and let $\{T_n\}$ be a sequence such that

$$
\Lambda(x,\vec{\delta},\vec{\xi}) = \lim_{n\to\infty} \frac{1}{T_n} E_x^{\vec{\delta},\vec{\xi}}\left[\sum_{t=0}^{T_n-1}[c(x_t,a_t) - I(\xi_t||\pi(\cdot|x_t,a_t))]\right].
$$

Then, given $M > 0$, there exists $N > 0$ such that, for $n > N$,

$$
\begin{aligned}
M &\leq \frac{1}{T_n} E_x^{\vec{\delta},\vec{\xi}} \sum_{t=0}^{T_n-1}[c(x_t,a_t) - I(\xi_t||\pi(\cdot|x_t,a_t))] \\
&\leq \frac{1}{T_n}\Lambda_{0,T_{n-1}}(x,\vec{\delta}) \\
&\leq \frac{1}{T_n} J_{0,T_{n-1}}(x,\vec{\delta}),
\end{aligned} \tag{3.5}
$$

where we have used part (a) of the theorem. Therefore, letting $n \to \infty$ in (3.5), and using part (b), we obtain

$$M \leq J(x, \vec{\delta}).$$

Since M was chosen arbitrarily, this inequality implies that $J(x, \vec{\delta}) = \infty$, which is a contradiction. Thus $\Lambda(x, \vec{\delta}, \vec{\xi}) < \infty$. Then, using essentially the same kind of arguments as in (3.5), (3.4) follows. □

Lemma 3.2.

(a) *For each* $\beta \in (0, 1)$ *and* $x \in S$,

$$0 \leq V_\beta(x) < \infty$$

and

$$\limsup_{\beta \to 1} (1 - \beta) V_\beta(x) \leq \rho.$$

(b) *The upper value function* V_β *is the minimal nonnegative solution of the Isaacs equation*

$$V_\beta(x) = \inf_{a \in A(x)} \sup_{\mu \in \Delta(x,a)} \left[c(x, a) - I(\mu || \pi(\cdot|x, a)) + \beta \int V_\beta \, d\mu \right], \tag{3.6}$$

where $\Delta(x, a) = \{\mu \in P(S) : I(\mu || \pi(\cdot||x, a)) < \infty\}$.

(c) *The stationary strategies* f^*_β *and* ξ^*, *with*

$$f^*_\beta(x) \in \arg\min \left\{ e^{c(x,a)} \int e^{\beta V_\beta(y)} \pi(dy|x, a) \right\}$$

and

$$\xi^*(x''|x, a) = \frac{e^{\beta V_\beta(x'')} \pi(x''|x, a)}{\int e^{\beta V_\beta(y)} \pi(dy|x, a)}$$

are optimal, whenever $\xi^* \in Q(f_\beta)$.

Proof. Let $\bar{f} \in \mathbf{F}$ be as in Assumption A.2(a), and let $x \in S$ be arbitrary, but fixed. Now we choose $\vec{\xi} \in Q(\bar{f})$ such that $V_\beta(x, \bar{f}, \vec{\xi}) \geq 0$. Then, using a well-known Tauberian theorem (see, e.g., [SF]),

$$\begin{aligned} \limsup_{\beta \to 1} (1 - \beta) V_\beta(x, \bar{f}, \vec{\xi}) &\leq \Lambda(x, \bar{f}, \vec{\xi}) \\ &\leq J(x, \bar{f}) \\ &= \rho, \end{aligned}$$

where we have used Theorem 3.1. Part (a) follows in a straightforward manner.

(b) Let $\beta \in (0, 1)$ be fixed. For each function $\psi \colon S \to \mathbb{R}$ define the operator

$$T_\beta \psi(x) := \min_{a \in A(x)} \left\{ c(x, a) + \log \int e^{\beta \psi(y)} \pi(dy|x, a) \right\}.$$

It is easy to see that T_β is monotone, i.e., if $\psi \geq \mu$, then $T_\beta \psi \geq T_\beta \mu$. Let $\psi_0 \equiv 0$ and define

$$\psi_{n+1} := T_\beta \psi_n.$$

Since $\{\psi_n\}$ is a nondecreasing sequence, there exists a nonnegative function ψ such that $\psi_n \uparrow \psi$. Then following analogous arguments to those used by Hernández-Lerma and Lasserre [HL2, Theorem 3.1], together with Lemma A.1, it can be seen that ψ satisfies the Isaacs equation (3.6). Further, ψ is the minimal nonnegative solution to this equation. Now we prove that $\psi = V_\beta$. Let f be a stationary policy such that

$$f(x) \in \arg\min_{a \in A(x)} \left\{ c(x,a) + \log \int e^{\beta\psi(y)} \pi(dy|x,a) \right\}.$$

Then, for any admissible policy $\vec{\xi} \in Q(f)$ for the second player and any $n \geq 1$,

$$\begin{aligned} \psi(x) &\geq \sum_{t=0}^{n} E_x^{f,\vec{\xi}} \beta^t [c(x_t, a_t) - I(\xi_t || \pi(\cdot|x_t, a_t))] + \beta^{n+1} E_x^{f,\vec{\xi}} \psi(x_{t+1}) \\ &\geq \sum_{t=0}^{n} E_x^{f,\vec{\xi}} \beta^t [c(x_t, u_t) - I(\xi_t || \pi(\cdot|x_t, u_t))]. \end{aligned}$$

Letting $n \to \infty$, this implies that

$$\psi(x) \geq V_\beta(x, f, \vec{\xi}).$$

Since $\vec{\xi}$ was chosen arbitrarily, we have that

$$\begin{aligned} \psi(x) &\geq \sup_{\vec{\xi} \in Q(f)} V_\beta(x, f, \vec{\xi}) \\ &\geq V_\beta(x). \end{aligned} \tag{3.7}$$

To prove the reverse inequality, we use the fact that the function ψ_n is the value function of the n-stage problem with terminal cost zero (see [HL1]). The proof of this fact is standard and is left to the reader. Thus, for each $x \in S$,

$$\psi_n(x) = \inf_{\vec{\delta} \in \Delta_M} \sup_{\vec{\xi} \in Q(\vec{\delta})} E_x^{\vec{\delta},\vec{\xi}} \sum_{t=0}^{n-1} \beta^t [c(x_t, a_t) - I(\xi_t || \pi(\cdot|x_t, a_t))].$$

Then, for any policy $\vec{\delta}$, $x \in S$ and $n = 1, 2, \ldots,$

$$\begin{aligned} \psi_n(x) &\leq \sup_{\vec{\xi} \in Q(\vec{\delta})} E_x^{\vec{\delta},\vec{\xi}} \sum_{t=0}^{n-1} \beta^t [c(x_t, a_t) - I(\xi_t || \pi(\cdot|x_t, a_t))] \\ &\leq \sup_{\vec{\xi} \in Q(\vec{\delta})} E_x^{\vec{\delta},\vec{\xi}} \sum_{t=0}^{\infty} \beta^t [c(x_t, a_t) - I(\xi_t || \pi(\cdot|x_t, a_t))]. \end{aligned}$$

Therefore,

$$\psi(x) \le \sup_{\vec{\xi} \in Q(\vec{\delta})} V_\beta(x, \vec{\delta}, \vec{\xi})$$

and then

$$\psi \le V_\beta(x).$$

Together with (3.7), this completes the proof of (b). The rest of the lemma follows immediately from standard dynamic programming arguments and Lemma A.1. □

Lemma 3.3. *There exists a finite set G and $\beta_0 \in (0, 1)$ such that, for each $\beta \in (\beta_0, 1)$, with β_0 as in Lemma* 3.2, *and $x \in S$,*

$$V_\beta(x) - V_\beta(x_\beta) \ge 0$$

for some $x_\beta \in G$. In addition, for any sequence $\{\beta_n\}$ converging to 1, *there exists a subsequence $\{\beta_{n_k}\}$ such that the sequence $\{x_{\beta_{n_k}}\}$ is constant.*

The proof of this lemma is a slight variation of the one given by Cavazos-Cadena [C] (see also [CS]), and we omit it.

4. Risk-Sensitive Optimal Controls

In this section we present our main result (see [HL3] for similar results in the risk-neutral case).

Theorem 4.1. *Under Assumptions A.*1 *and A.*2, *there exist a number ρ^* and a* (*possibly extended*) *function W on S such that, for all $x \in S$,*

$$e^{\rho^* + W(x)} \ge \inf_{a \in A(x)} \left\{ e^{c(x,a)} \int e^{W(y)} \pi(dy|x, a) \right\}$$

and the set $H := \{x \in S : W(x) \text{ is finite}\}$ is not empty. Moreover, there exists an optimal control $f^* \in \mathbf{F}$ whenever the initial state belongs to H, and

$$\rho^* = J(x, f^*)$$

for all $x \in H$.

Proof. Let $\{\beta_n\}$ be a sequence in $(0, 1)$ converging to 1, and take a subsequence (also denoted by $\{\beta_n\}$) as in Lemma 3.3, labeling by e the common value of the sequence

$\{x_{\beta_n}\}$. Following a standard approach, we define $\rho_n := (1-\beta_n)V_{\beta_n}(e)$, $\quad W_n(x) := V_{\beta_n}(x) - V_{\beta_n}(e)$, and $W_\beta(x) := V_\beta(x) - V_\beta(e)$, and rewrite (3.6), using Lemma A.1, as

$$e^{\rho_n + W_n(x)} = \min_{a\in A(x)} \left\{ e^{c(x,a)} \int e^{\beta_n W_n(y)} \pi(dy|x,a) \right\}. \tag{4.1}$$

We define $\rho^* := \limsup_n \rho_n$ and $W(x) := \liminf_n W_n(x)$; then, taking the $\liminf_n$ on both sides of (4.1), and using Fatou's lemma and Assumption A.1, we conclude that

$$\begin{aligned} e^{\rho^* + W(x)} &\geq \liminf_n \min_{a\in A(x)} \left\{ e^{c(x,a)} \int e^{\beta_n W_n(y)} \pi(dy|x,a) \right\} \\ &\geq \min_{a\in A(x)} \left\{ e^{c(x,a)} \int e^{W(y)} \pi(dy|x,a) \right\}. \end{aligned} \tag{4.2}$$

On the other hand, from the definition of function W, it follows that at least e belongs to H. Now, let $f^* \in \mathbf{F}$ achieve the minimum on the right-hand side of (4.2).

It remains to prove that f^* is optimal. First, we prove that, for any control $\vec{\delta} \in \Delta_M$, with $J(x,\vec{\delta}) \leq \rho$, and $x \in S$,

$$\rho^* \leq J(x,\vec{\delta}). \tag{4.3}$$

Let $x \in S$. Then, by Lemma 3.3, for each $\beta \in (\beta_0, 1)$,

$$\begin{aligned} (1-\beta)V_\beta(x) &= (1-\beta)W_\beta(x) + (1-\beta)V_\beta(e) \\ &\geq (1-\beta)V_\beta(e), \end{aligned}$$

which implies

$$\rho^* \leq \limsup_{\beta\to 1} (1-\beta)V_\beta(x). \tag{4.4}$$

Now let $\vec{\delta} \in \Delta_M$ such that $J(x,\vec{\delta}) \leq \rho$, and choose $\vec{\xi} \in Q(\delta)$ such that $V_\beta(x,\vec{\delta},\vec{\xi}) \geq 0$. Then by a well-known Tauberian theorem and (3.4), we obtain

$$\begin{aligned} \limsup_{\beta\to 1} (1-\beta)V_\beta(x,\vec{\delta},\vec{\xi}) &\leq \Lambda(x,\vec{\delta},\vec{\xi}) \\ &\leq J(x,\vec{\delta}). \end{aligned}$$

Therefore, it follows that

$$\limsup_{\beta\to 1} (1-\beta)V_\beta(x) \leq J(x),$$

which together with (4.4) implies (4.3). We prove now that $\rho^* \geq J(x, f^*)$ whenever $x \in H$. From (4.2), we have that, for any $x \in H$,

$$\begin{aligned} E_x^{f^*} \exp\left[\sum_{t=0}^{T-1} c(x_t,a_t)\right] &\leq e^{\rho^* T} E_x^{f^*}\left[\prod_{t=0}^{T-1} \frac{e^{W(x_t)}}{\int e^{W(y)}\pi(dy|x_t,a_t)}\right] \\ &\leq e^{\rho^* T} \cdot \frac{e^{W(x)}}{\inf_{x\in H} \int e^{W(y)} \pi(dy|x, f^*(x))}, \end{aligned}$$

where the last inequality follows from standard properties of conditional expectations and the Markov property.

Therefore,

$$J(x, f^*) \leq \rho^*. \tag{4.5}$$

Then (4.5) and (4.3) imply the optimality of f^*. □

Appendix

The next lemma establishes a variational formula for the logarithmic moment-generating function. We refer to Proposition 4.5.1 of [DE] for its proof.

Lemma A.1. *Let ψ be a real-valued function defined on S bounded from below, and let ν be a probability measure on $P(S)$. Then*

$$\log \int e^{\psi}\, d\nu = \sup_{\mu \in \Delta(\nu)} \left\{ \int \psi\, d\mu - I(\mu || \nu) \right\}, \tag{A.1}$$

where $\Delta(\nu) := \{\mu \in P(S) : I(\mu||\nu) < \infty\}$. Morever, the supremum on the right-hand side of (A.1) *is attained at μ^* defined by*

$$\mu^*(x) := \frac{e^{\psi(x)}\nu(x)}{\int e^{\psi}\, d\nu}, \qquad x \in S,$$

whenever $\int e^{\psi}\, d\nu$ is finite.

References

[ABF+] Arapostathis A, Borkar VS, Fernández-Gaucherand E, Ghosh MK, Marcus SI (1993) Discrete-time controlled Markov processes with average cost criterion: a survey. SIAM J Control Optim 31:282–344

[B] Borkar VS (1984) On minimum cost per unit of time control of Markov chains. SIAM J Control and Optim 22:965–978

[BJ] Baras JS, James MR (1997) Robust and risk sensitive output feedback control for finite state machines and hidden Markov models. J Math Systems Estimation Control 7:371–374

[C] Cavazos-Cadena R (1989) Weak conditions for the existence of optimal stationary policies in average Markov decision chains with unbounded costs. Kybernetika 25:145–156

[CS] Cavazos-Cadena R, Sennott LI (1992) Comparing recent assumptions for the existence of average optimal stationary policies. Oper Res Lett 11:33–37

[DE] Dupuis P, Ellis RS (1997) A Weak Convergence Approach to the Theory of Large Deviations. Wiley, New York

[DMR] Dai Pra P, Meneghini L, Runggaldier WJ (preprint) Some connections between stochastic control and dynamic games

[DS] Deuschel J-D, Stroock DW (1989) Large Deviations, Academic Press, Boston

[FGM] Fernández-Gaucherand E, Marcus SI (1994) Risk sensitive optimal control of hidden Markov models: a case study. Proc 33rd IEEE Conf on Decision and Control, pp. 1657-1662

[FH] Fleming WH, Hernández-Hernández D (1997) Risk sensitive control of finite state machines on an infinite horizon I. SIAM J Control Optim 35:1790–1810

[FM1] Fleming WH, McEneaney WM (1992) Risk-Sensitive and Differential Games. Lecture Notes in Control and Information Science No 184. Springer-Verlag, Berlin, pp. 185–197

[FM2] Fleming WH, McEneaney WM (1995) Risk-sensitive control on an infinite horizon. SIAM J Control Optim 33:1881–1915

[HL1] Hernández-Lerma O, Lasserre JB (1996) Discrete-Time Markov Control Processes. Basic Optimality Criteria. Springer-Verlag, New York

[HL2] Hernández-Lerma O, Lasserre JB (1990) Average cost optimal policies for Markov control processes with Borel state space and unbounded costs. Systems Control Lett 15:349–356

[HL3] Hernández-Lerma O, Lasserre JB (1994) Weak conditions for average optimality in Markov control processes. Systems Control Lett 22:287–291.

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

[S1] Sennott LI (1986) A new condition for the existence of optimal stationary policies in average cost Markov decision processes. Oper Res Lett 5:17–23

[S2] Sennott LI (1989) Average cost optimal stationary policies with unbounded costs. Oper Res 37:626–633

[SF] Sznajder R, Filar JA (1992) Some comments on a theorem of Hardy and Littlewood. J Optim Theory Appl 75:201-208

[W] Whittle P (1990) Risk-Sensitive Optimal Control. Wiley, New York

Accepted 1 *October* 1997