ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/268064711

Risk Sensitive Markov Decision Processes

Article · January 1997

DOI: 10.1007/978-1-4612-4120-1_14

CITATIONS
28

READS
149

5 authors, including:

Daniel Hernández-Hernández
Centro de Investigación en Matemáticas (CIMAT)
70 PUBLICATIONS 756 CITATIONS

SEE PROFILE

All content following this page was uploaded by Daniel Hernández-Hernández on 04 May 2016.

The user has requested enhancement of the downloaded file.

Risk Sensitive Markov Decision Processes

Steven I. Marcus[1]
Emmanuel Fernández-Gaucherand[2]
Daniel Hernández-Hernández[3]
Stefano Coraluppi[1]
Pedram Fard[1]

1 Introduction

Risk-sensitive control is an area of significant current interest in stochastic control theory. It is a generalization of the classical, risk-neutral approach, whereby we seek to minimize an exponential of the sum of costs that depends not only on the expected cost, but on higher order moments as well.

Research effort has been directed towards establishing results that parallel those already available in the risk-neutral setting, as well as towards exploring the connections of risk-sensitive control to robust control and differential games. This paper summarizes some contributions to the first of these objectives.

For linear systems and exponential of the sum of quadratic costs, the problem has been studied by [26] in the fully observed setting. Extensions to the partially observed setting are due to [3] and [34]. A somewhat surprising result is that the conditional distribution of the state given past observations does not constitute an information state. The equivalence, in the large risk limit, to a differential game arising in H^∞ control is due to [12] and [20]. Nonlinear systems have been studied in [18, 31, 34]. Partially observed nonlinear systems are treated in [27], where the appropriate information state and dynamic programming equations are presented.

In parallel to the work in the control community, there has been a body of research on risk-sensitive Markov Decision Processes (MDPs) –

[1] Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, Maryland 20742. Research supported in part by the National Science Foundation under grant EEC 9402384.

[2] Systems & Industrial Engineering Department, University of Arizona, Tucson, Arizona 85721. Research supported in part by a grant from the University of Arizona Foundation and the Office of the Vice President for Research; and in part by the National Science Foundation under grant NSF-INT 9201430.

[3] Department of Mathematics, CINVESTAV-IPN, Apartado postal 14-740 Mexico D.F. 07000, MEXICO.

i.e., discrete-time problems with a finite or countable state space. An early formulation of the risk-sensitive MDP problem is due to [25]. Discounted cost problems are studied in [8]; a surprising result is that, for the infinite horizon problem with discounted costs, the optimal policy is not stationary in general. An information state and dynamic programming equations for the partially observed problem are introduced in [2]. Structural results for the value function in the partially observed setting are provided in [15]. The average cost problem has been studied in [16], [17], [22], [23], [25].

The nonstationarity of the discounted problem has motivated an alternative generalization of the risk-neutral formulation, which preserves the stationarity of the optimal control law. The formulation is developed in [28], and has been studied recently in the linear systems context in [21]. Further results and algorithms for the discounted problem, including the problem with partial observations, are presented in [9].

2 The Risk Sensitive MDP Model

In this section we give the general idea of the problem formulation; specific assumptions will be stated in later sections. We restrict attention to discrete-time stochastic dynamical systems with finite or countable state and observation spaces, and finite or compact action (or control) set. For this class of systems, we can employ an MDP (or controlled Markov chain) description. This is given by $M = (X, Y, U, \{P(u), u \in U\}, \{Q(u), u \in U\})$, where X is the state space, Y is the output space, U is the set of controls. X_t, Y_t, and U_t denote the state, output, and control at time t. $P(u)$ and $Q(u)$ are the state transition matrix and the output matrix, respectively, for $u \in U$. More precisely, $p_{i,j}(u) := pr(X_{t+1} = j | X_t = i, U_t = u)$ (also denoted $P(j|i,u)$) and $q_{i,y}(u) := pr(Y_t = y | X_t = i, U_{t-1} = u)$; in addition, we define the matrix $\overline{Q}(y,u) := diag(q_{i,y})(u)$.

A policy or control law is a sequence of mappings $\pi = (\pi_0, \pi_1, \ldots)$ such that $u_k = \pi_k(Y^k), Y^k = (Y_1, \ldots, Y_k)$. Let us denote by $\mathbf{\Pi}$ the set of admissible policies. Traditionally, an additive cost structure has been employed, of the form

$$J^0(\pi) = E^\pi[C_M],$$

where $c(i,u)$ is the cost per stage and $C_M = \sum_{t=0}^{M-1} c(X_t, U_t)$.

The *finite horizon risk-sensitive control* problem is to find a policy π to minimize

$$J^\gamma(\pi) := \gamma \log \mathbb{E}^\pi \left[exp\left(\gamma^{-1} \cdot \mathcal{C}_M\right)\right]$$

where $\gamma^{-1} \neq 0$ is the *risk factor;* in this paper we will consider the *risk-averse* case in which $\gamma > 0$. Notice that, to first order in γ^{-1},

$$J^{\gamma}(\pi) \simeq \mathbb{E}^{\pi}\big[\mathcal{C}_M\big] + \frac{1}{2\gamma} Var^{\pi}\big[\mathcal{C}_M\big].$$

The minimization of $J^{\gamma}(\pi)$ is equivalent to the minimization of

$$\overline{J}^{\gamma}(\pi) := \mathbb{E}^{\pi}\big[exp\big(\gamma^{-1}\cdot\mathcal{C}_M\big)\big].$$

We will also discuss the *average cost risk-sensitive control* problem, in which one seeks to minimize

$$J_a^{\gamma}(\pi) = \limsup_{T\to\infty} \frac{\gamma}{T}\log \mathbb{E}^{\pi}\exp\{\frac{1}{\gamma}\sum_{t=0}^{T-1} c(X_t,U_t)\}.$$

3 Complete State Observations

In this section, we assume that $Y_t = X_t$ – i.e., that we have complete observations of the state.

3.1 The Finite Horizon Case

This problem was first considered in [25]. Define the value function by

$$S_{k,M}^{\gamma}(x) := \min_{\pi} \mathbb{E}^{\pi}\big[exp\big(\gamma^{-1}\cdot\mathcal{C}_{k,M}\big)|X_k = x\big].$$

where $\mathcal{C}_{k,M} = \sum_{t=k}^{M-1} c(X_t,U_t)$. The value function satisfies the dynamic programming recursion:

$$\begin{array}{rcl} S_{N,N}^{\gamma} & = & 1 \\ S_{k,N}^{\gamma} & = & \min_{u\in U}\{\mathcal{D}(u)S_{k+1,N}^{\gamma}\} \end{array}$$

where the minimum is taken separately for each component of the vector equation and $[\mathcal{D}(u)]_{i,j} := p_{i,j}(u)\cdot exp(\gamma^{-1}c(i,u))$ is the "disutility contribution matrix". There exists a Markov policy (i.e., a policy such that π_k depends only on the state X_k) that is optimal.

3.2 Infinite Horizon, Average Cost

For the finite state case, this problem has been studied in [25] and, more recently, in [16]; the latter paper also discusses the relation to robust control problems. We will present results for the countable state case, following [22, 23]. As in [22], we assume first that U is Borel space; for $x \in X$, $U(x)$ is the set of admissible actions. In addition, we will assume:

Assumption A.1.

(i) For each $x \in X, U(x)$ is a compact subset of U.

(ii) The cost function c is nonnegative, continuous and bounded.

(iii) For all $x, z \in X$, the function $u \mapsto P(z|x,u)$ is continuous on $U(x)$.

The average cost risk sensitive optimal control problem is to find a policy $\pi^* \in \mathbf{\Pi}$ that minimizes $J_a^\gamma(\pi)$. Define

$$\Lambda := \inf_{\pi \in \mathbf{\Pi}} J_a^\gamma(\pi).$$

The main objective is to find sufficient conditions to ensure the existence of a stationary optimal policy.

Verification Theorem.

Theorem 3.1 [22]. Suppose that there exist a number λ and a bounded function $W : X \to \mathbb{R}$ such that

$$e^{\lambda + W(x)} = \min_{u \in U(x)} \{ e^{\gamma^{-1} c(x,u)} \sum_{z \in X} e^{W(z)} P(z|x,u) \}. \tag{3.1}$$

Then

$$\lambda\gamma \leq J_a^\gamma(\pi) \text{ for all } \pi \in \mathbf{\Pi}.$$

Further, if π^* is a stationary policy, with $\pi^*(x)$ achieving the minimum on the r.h.s. of (3.1) for each $x \in X$, then π^* is optimal, and

$$\lambda = \lim_{T \to \infty} \frac{1}{T} \log \mathbb{E}^{\pi^*} \exp \{ \frac{1}{\gamma} \sum_{t=0}^{T-1} c(X_t, U_t) \}.$$

Remark. Notice that the right hand side of (3.1) looks like a moment generating function; we will use this fact later.

Existence of Solutions.

We turn to the question of existence of a solution to the dynamic programming equation (3.1). The main result is the following:

Theorem 3.2 [22]. For each $e \in X$ and $\pi \in \mathbf{\Pi}$ define

$$\tau_e := \min \{t > 0 : x_t = e\}.$$

If there exist $e \in X$ and $C > 0$ such that

$$\mathbb{E}^{\pi}(\tau_e | X_0 = x) < C \tag{3.2}$$

for all $\pi \in \mathbf{\Pi}$ and $x \in X$, then there exists a solution (λ, W) to the dynamic programming equation (3.1), with W bounded.

Remarks. (i) Similar results are proved under weaker conditions in [23], and are briefly discussed below.

(ii) One might expect that this theorem can be proved by using the "vanishing discount" approach that has been so successful in the risk neutral case (see, e.g., [1] and the references therein); in this approach, one solves the corresponding discounted problem and obtains the solution of the average cost problem as a limit of discounted problems as the discount factor approaches 1. However, in the risk sensitive case, the optimal policies for the "corresponding" discounted problem with cost

$$\mathbb{E}^{\pi}\left[exp(\gamma^{-1}\sum_{t=0}^{\infty}\beta^t c(X_t, U_t))\right] \tag{3.3}$$

are *not stationary* [8]! An intuitive explanation is that the decision maker appears less risk averse, by a factor β, from step to step, approaching risk-neutrality as $t \to \infty$. An alternative approach, sketched here, employs instead a sequence of discounted dynamic games [16, 22].

This approach depends in a fundamental way on a duality result. Let $P(X)$ be the set of probability vectors on X, i.e.

$$P(X) = \{\mu = (\mu^0, \mu^1, \ldots) : \mu^i \geq 0, \sum_{i \in X} \mu^i = 1\}.$$

Fix $\nu \in P(X)$, and define the relative entropy function $I(\cdot||\nu) : P(X) \to \mathbb{R} \cup \{+\infty\}$ by

$$I(\mu||\nu) = \begin{cases} \sum\limits_{x \in X} log(r(x))\mu(x) & \text{if } \mu << \nu \\ +\infty & \text{otherwise} \end{cases}$$

where

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

The next lemma establishes, using a Legendre-type transformation, the duality relationship between the relative entropy function and the logarithmic moment generating function.

Lemma 3.3 [10, Proposition II.4.2]. Let ψ be a bounded function defined on X, and let $\nu \in P(X)$. Then,

$$\log \sum_{z \in X} e^{\psi(z)} \nu(z) = \sup_{\mu \in P(X)} \{ \sum_{z \in X} \psi(z)\mu(z) - I(\mu||\nu) \};$$

the supremum is attained at the unique probability measure μ^* defined by

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

Using Lemma 3.3, we rewrite equation (3.1) as

$$\lambda + W(x) \;=\; \min_{u \in U(x)} \sup_{\mu \in P(X)} \{ \sum W\mu + \frac{1}{\gamma} c(x,u) - I(\mu||P(\cdot|x,u)) \} \quad (3.4)$$

This equation corresponds to the Isaacs equation associated with a stochastic dynamic game with average cost per unit time criterion (see [11, 22]).

Theorem 3.1 can then be proved via the vanishing discount approach, by first considering the corresponding infinite horizon discounted cost stochastic dynamic game. Let W_β be the upper value function of this game. Then, once we find a uniform bound for a "differential" discounted value function, i.e. $h_\beta(x) := W_\beta(x) - W_\beta(e)$, with e as in (3.2), the theorem follows by letting $\beta \to 1$.

First we introduce the infinite horizon discounted cost dynamic game.

Stochastic dynamic game. Let X be the state space, U be the control set for Player 1 (minimizer), and $P(X)$ be the control set for Player 2 (maximizer). The reward function is $(x, u, \mu) \mapsto \frac{1}{\gamma} c(x,u) - I(\mu||P(\cdot|x,u))$.

The evolution of the system is as follows (c.f. [16] [22]). At each time $t \in \{0, 1, \ldots\}$ the state of the system is observed, say $X_t = x \in X$. Then, a control $U_t \in U(x)$ is chosen for Player 1, and $\mu_t \in P(X)$ is chosen for Player 2. Then, a reward $\frac{1}{\gamma} c(X_t, U_t) - I(\mu_t||P(\cdot|X_t, U_t))$ is earned, and the state of the system moves to the state X_{t+1} according to the probability distribution μ_t.

Strategies. For each $t \geq 0$, let N_t and K_t be the set of feasible histories up to time t for Player 1 and Player 2, respectively. That is, $N_0 = S$ and $N_t =$

$(S\times P(S))^t \times S$, while $K_0 = K$ and $K_t = K^t \times K$, where $K = \{(x,u) :\ u \in U(x), x \in X\}$. Generic elements of N_t and K_t are vectors of the form $n_t = (X_0, \mu_0, \ldots, X_{t-1}, \mu_{t-1}, X_t)$ and $K_t = (X_0, U_0, \ldots, X_{t-1}, U_{t-1}, X_t, U_t)$, respectively. A non-randomized strategy for Player 1 is a sequence $\pi = \{\pi_t\}$ of functions π_t from N_t to U, such that $\pi_t(n_t) \in U(X_t)$ for all $n_t \in N_t$. We say that π is stationary if, for all $t \geq 0, \pi_t$ depends only on the current state X_t, and π_t is independent of t. A non-randomized strategy for Player 2 is a sequence $\xi = \{\xi_t\}$ of functions ξ_t from K_t to $P(X)$. Stationarity of ξ is defined similarly.

Given the initial state $x \in X$, let $P_x^{\pi,\xi})$ be the probability induced by the strategies π, ξ , and $\mathbb{E}_x^{\pi,\xi}$ the corresponding expectation operator. Equation (3.3) corresponds to the dynamic progamming (Isaacs) equation of the stochastic dynamic game described above with average cost optimality criterion, defined for each $x \in X, \pi, \xi$ as

$$\Lambda(x,\pi,\xi) := \limsup_{T\to\infty} \mathbb{E}_x^{\pi,\xi} \frac{1}{T} \sum_{t=0}^{T-1} \left[\frac{1}{\gamma} c(X_t, U_t) - I(\xi_t || P(\cdot|X_t, U_t)) \right].$$

The corresponding discounted games are defined via the cost functionals

$$J_\beta(x,\pi,\xi) = \mathbb{E}_x^{\pi,\xi} \sum_{t=0}^{\infty} \beta^t \left[\frac{1}{\gamma} c(x_t, a_t) - I(\xi_t || P(\cdot|X_t, U_t)) \right],$$

where $\beta \in (0,1)$ is the discount factor.

Definition 3.4. When there exist a pair of strategies (π^*, ξ^*) such that

$$J_\beta(x,\pi^*,\xi) \leq J_\beta(x,\pi^*,\xi^*) \leq J_\beta(x,\pi,\xi^*)$$

for all π, ξ, the value $W_\beta(x) = J_\beta(x,\pi^*,\xi^*)$ is called the value of the game, and (π^*, ξ^*) are referred to as *optimal strategies*.

Lemma 3.5 [22]. There is a unique bounded solution to the Isaacs equation

$$W_\beta(x) = \min_{u\in U(x)} \sup_{\mu \in P(X)} \{ \sum_z \beta W_\beta(z) \mu(z) + \frac{1}{\gamma} c(x,u) - I(\mu || P(\cdot|x,u)) \}, \tag{3.5}$$

and it is the value function of the discounted cost stochastic dynamic game. Moreover, stationary strategies are optimal.

Remark. Note that, by Lemma 3.3, equation (3.5) can be rewritten as

$$e^{W_\beta(x)} = \min_{u\in U(x)} \{ e^{\frac{1}{\gamma} c(x,u)} \sum_z e^{\beta W_\beta(z)} P(z|x,u) \}. \tag{3.6}$$

This optimality equation has been studied by Eagle [11] in the context of a particular type of risk-sensitive discounted Markov decision process. Chung and Sobel [8] (see also the references therein), study risk-sensitive discounted Markov decision processes with a very different optimality equation, which results in *nonstationary* optimal policies (see Section 5 below).

Now, to employ the vanishing discount approach, we define $h_\beta(x) := W_\beta(x) - W_\beta(e)$, with e as in (3.2), and write (3.6) as

$$e^{(1-\beta)W_\beta(e)} \cdot e^{h_\beta(x)} = \min_{u \in U(x)} e^{\frac{1}{\gamma}c(x,u)} \sum e^{\beta h_\beta(z)} P(z|x,u)\}. \tag{3.7}$$

Sketch of proof of Theorem 3.1 [22]. It is first proved (using the assumption (3.2) and the boundedness of c) that $(1-\beta)W_\beta(e)$ and h_β are uniformly bounded. Let $\beta_n \uparrow 1$ be given. Then, boundedness of $(1-\beta)W_\beta(e)$ and h_β imply that, by a suitable diagonalization, we may pick a subsequence $\{\beta_n\}$ (denoting it again by $\{\beta_n\}$) along which $h_{\beta_n}(x), x \in X$, and $(1-\beta)W_\beta(e)$ converge to some limits $W(x)$ and λ, respectively. Thus, the theorem follows from (3.7) and an application of the Dominated Convergence Theorem. ■

In [23], this problem is studied under considerably weaker hypotheses, similar to those used in previous literature for the risk-neutral average cost criterion [5]-[7].

Assumption A.2.

(i) For each $x, z \in X$, the mapping $u \to P(z|x,u)$, with $u \in U(x)$ is lower semi-continuous.

(ii) For each $x \in X, U(x)$ is a compact subset of U.

Define

$$J_a^\gamma(x,\pi) = \limsup_{T\to\infty} \frac{\gamma}{T} \log \mathbb{E}^\pi[\exp\{\frac{1}{\gamma}\sum_{t=0}^{T-1} c(X_t,U_t)\}|X^0 = x].$$

Assumption A.3 (a) There exists a stationary policy $\bar{\pi} \in \mathbf{\Pi}$ such that

$$\rho := J_a^\gamma(x,\bar{\pi})$$

is finite and independent of x.

(b)

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

The following theorem, which presents a dynamic programming *inequality*, is proved via the dynamic stochastic game and vanishing discount approach discussed above.

Theorem 3.6 [23]. Under Assumptions A.2 and A.3, there exist a number ρ^* and a (possibly extended) function W on X such that for all $x \in X$

$$e^{\rho^*+W(x)} \geq \inf_{u \in U(x)} \{e^{c(x,u)} \sum e^{W(z)} P(z|x,u)\}$$

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

$$\rho^* = J_a^\gamma(x, \pi^*)$$

for all $x \in H$.

4 Partial State Observations

In this section, we discuss risk sensitive Markov decision processes with partial state observations, also know as *hidden Markov models*. We assume throughout that X, U, and Y are finite with cardinalities N_X, N_U, and N_Y, respectively.

4.1 The Finite Horizon Case

As for the risk-neutral case [1], [4], [29], an equivalent stochastic optimal control problem can be formulated in terms of *information states* and *separated policies*. Here we follow the work of Baras, Elliott, and James [2], [27]. Let $\mathcal{Y}_t$ be the filtration generated by the available observations up to time t, and let $\mathcal{G}_t$ be the filtration generated by the sequence of states and observations up to that time. Then the probability measure induced by a policy π is equivalent to a canonical distribution $\mathcal{P}^\dagger$, under which $\{Y_t\}$ is independently and identically distributed (i.i.d), uniformly distributed, independent of $\{X_t\}$, and $\{X_t\}$ is a controlled Markov chain with transition matrix $P(u)$. Also,

$$\frac{d\mathcal{P}^\pi}{d\mathcal{P}^\dagger}|_{\mathcal{G}_t} = \lambda_t^\pi := N_Y^t \cdot \Pi_{k=1}^t q_{X_k,Y_k}(U_{k-1}).$$

The cost incurred by using the policy π is given by

$$\overline{J}^\gamma(\pi) = \mathbb{E}^\dagger[\lambda_M^\pi exp(\gamma^{-1} \cdot \mathcal{C}_M)]$$

Following [2], [27], the information state is given by

$$\sigma_t^\gamma(i) := \mathbb{E}^\dagger\big[I[X_t = i]exp\big(\gamma^{-1}\cdot\mathcal{C}_t\big)\cdot\lambda_t^\pi \mid \mathcal{Y}_t\big],$$

where $I[A]$ is the indicator function of the event A, and $\sigma_0^\gamma(i) = p_0$, where p_0 is the initial distribution of the state and is assumed to be known. With this definition of information state, similar results as in the risk-neutral case can be obtained. In particular, one obtains a recursive updating formula for $\{\sigma_t^\gamma\}$, which is driven by the output (observation) path and evolves forward in time. Moreover, the value functions can be expressed in terms of the information state only, and dynamic programming equations give necessary and sufficient optimality conditions for *separated policies*, i.e., maps $\sigma_t^\gamma \mapsto \tilde{\pi}_t(\sigma_t^\gamma) \in U$; see [2], [27]. In particular we have that:

$$\overline{J}^\gamma(\pi) = \mathbb{E}^\dagger\big[\sum_{i=1}^{N_X}\sigma_M^\gamma(i)\big],$$

where $\{\sigma_M^\gamma\}$ is obtained under the action of policy π.

General Results.

The following lemma gives the recursions that govern the evolution of the information state.

Lemma 4.1 [2], [27]. The information state process $\{\sigma_t^\gamma\}$ is recursively computable as:

$$\sigma_{t+1}^\gamma = N_Y \cdot M(Y_{t+1}, U_t)\sigma_t^\gamma, \tag{4.1}$$

where

$$M(Y_{t+1}, U_t) := \overline{Q}(Y_{t+1}, U_t)\mathcal{D}^T(U_t),$$

$\overline{Q}(y,u) := diag(q_{i,y})(u)$, $\mathcal{D}$ is defined in Section 3.1, T denotes transpose.

Remark. Observe that as $\gamma^{-1} \to 0$, $\mathcal{D}(u) \to P(u)$ (elementwise). Therefore, (4.1) is the "natural" extrapolation of the (unnormalized) conditional probability distribution of the (unobservable) state, given the available observations, which is the standard risk-neutral information state [1], [4], [29].

Define value functions $J^\gamma(\cdot, M-k) : \mathbb{R}_+^{N_X} \to \mathbb{R}$, $k = 1, \ldots, M$, as follows:

$$J^\gamma(\sigma, M-k) := \min_{\pi_{M-k}\cdots\pi_{M-1}}\Big\{\mathbb{E}^\dagger\big\{\sum_{i=1}^{N_X}\sigma_M^\gamma(i) \mid \sigma_{M-k}^\gamma = \sigma\big\}\Big\}.$$

Lemma 4.2 [2]. The dynamic programming equations for the value functions are:

$$\begin{aligned} J^{\gamma}(\sigma, M) &= \sum_{i=1}^{N_X} \sigma(i); \\ J^{\gamma}(\sigma, M-k) &= \min_{u \in U} \big\{ \mathbb{E}^{\dagger}\big[J^{\gamma}(N_Y M(u, Y_{M-k+1}) \cdot \sigma, M-k+1) \big] \big\}, \\ k = 1, 2, \ldots, M. & \qquad (4.2) \end{aligned}$$

Furthermore, a separated policy $\pi^* = \{\pi_0^*, \ldots, \pi_{M-1}^*\}$ that attains the minimum in (4.2) is risk-sensitive optimal.

The following generalize similar structural results for the standard risk-neutral case [1], [4], [13], [29], [32].

Lemma 4.3 [15]. The value functions given by (4.2) are concave and piecewise linear functions of $\sigma \in \mathbb{R}_+^{N_X}$.

Lemma 4.4 [15]. Optimal separated policies $\{\pi_t^*\}$ are constant along rays through the origin, i.e., if $\sigma \in \mathbb{R}_+^{N_X}$ then $\pi_t^*(\sigma') = \pi_t^*(\sigma)$, for all $\sigma' = \alpha\sigma$, $\alpha \geq 0$.

An action $\overline{u} \in U$ is said to be a *resetting* action if there exists $j^* \in X$ such that $p_{i,j^*}(\overline{u}) = 1$, for all $i \in X$.

Using these results and a result of Lovejoy [30, Lemma 1], the next Theorem can be proved.

Theorem 4.5 [15]. Let $\overline{u} \in U$ be a resetting action. Then $CR_{\overline{u}}^k$, the region in which the control value $\overline{u}$ is optimal at time k, is a convex subset of $\mathbb{R}_+^{N_X}$.

A Case Study.

In [15], risk sensitive control of a popular benchmark problem is considered; much is known about this problem in the risk-neutral case. This is a two-state replacement problem which models failure-prone units in production/manufacturing systems, communication systems, etc. The underlying state of the unit can either be *working* ($X_t = 0$) or *failed* ($X_t = 1$), and the available actions are to *keep* ($U_t = 0$) the current unit or *replace* ($U_t = 1$) the unit by a new one. The cost function $(x, u) \mapsto c(x, u)$ is as follows: let $R > C > 0$, then $c(0,0) = 0$, $c(1,0) = C$, $c(x,1) = R$. The observations have probability $1/2 < q < 1$ of coinciding with the true state of the unit. The state transition matrices are given as:

$$P(0) = \begin{bmatrix} 1-\theta & \theta \\ 0 & 1 \end{bmatrix}; \quad P(1) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},$$

with $0 < \theta < 1$; see [13], [14], [33] for more details. With the above definitions, the matrices used to update the information state vector are given by:

$$M(y,0) = \begin{bmatrix} q_y(1-\theta) & 0 \\ (1-q_y)\theta & (1-q_y)e^{\gamma C} \end{bmatrix}; \quad M(y,1) = \begin{bmatrix} q_y e^{\gamma R} & q_y e^{\gamma R} \\ 0 & 0 \end{bmatrix}, \tag{4.3}$$

where $q_y := q(1-y) + (1-q)y$, $y = 0, 1$. For this case $\sigma = (\sigma(1), \sigma(2))^T \in \mathbb{R}^2_+$. Define the *replace* control region $CR^k_{replace}$ and the *keep* control region CR^k_{keep} in the obvious manner. The next result follows from (4.3), Lemma 4.4, and Theorem 4.5.

Lemma 4.6 [15]. For all decision epochs the *replace* control region is a (possibly empty) conic segment in $\mathbb{R}^2_+$.

The next result establishes an important *threshold* structural property of the optimal control policy. This is similar to well known results for the risk neutral case [13], [14], [30], [33].

Theorem 4.7 [15]. If $CR^k_{replace}$ is nonempty, then it includes the $\sigma(2)$-axis, i.e., $\mathbb{R}^2_+$ is partitioned by a line through the origin such that for values of $\sigma \in \mathbb{R}^2_+$ above the line it is optimal to *replace* the unit, and it is optimal to *keep* the unit otherwise.

Further structural results for this example, as well as results on limiting behavior for large and small risk factors, are presented in [15].

4.2 Infinite Horizon, Average Cost

Risk sensitive control of average cost, finite state, partially observed models has been studied in [17], using the approach discussed in Section 3.2 for the completely observed case. However, an information state is used in place of the state. In this paper, we will only consider the risk sensitive average cost version of the two-state replacement problem discussed above in Section 4.1 – i.e., with cost functional $J^\gamma_a(\pi)$; for the risk neutral case, this problem has been studied in detail in [13]. As suggested in [17], it is convenient for the infinite horizon problem to use a normalized information state

$$\rho^\gamma_t := \frac{\sigma^\gamma_t}{|\sigma^\gamma_t|} = \begin{bmatrix} 1-\alpha^\gamma_t \\ \alpha^\gamma_t \end{bmatrix}$$

where

$$|\sigma| := \sum_{j=1}^{N_Y} \sigma(j)$$

Thus α^γ_t can be used as the (one-dimensional) information state. Then

$$\rho^\gamma_{t+1} = \frac{M(Y_{t+1}, U_t)\rho^\gamma_t}{|M(Y_{t+1}, U_t)\rho^\gamma_t|}$$

or

$$\alpha_{t+1}^{\gamma} = f(\alpha_t^{\gamma}, Y_{t+1}, U_t),$$

where f is defined implicitly.

The dynamic programming equation corresponding to (3.1) is

$$e^{\lambda + W(\alpha)} = \min_{u \in U} \{ \mathbb{E}^{\dagger} \{ |M(y,u) \begin{bmatrix} 1-\alpha \\ \alpha \end{bmatrix} | e^{W(f(\alpha,y,u))} \}, \tag{4.4}$$

and we have the following Verification Theorem.

Theorem 4.8 [24]. Let (λ, W) be a solution of equation (4.4), with W bounded. Then the separated stationary policy π^*, with $\pi^*(\alpha)$ achieving the minimum in the r.h.s. of (4.4) is optimal and $\lambda\gamma$ is the optimal average cost.

In order to study the existence of solutions to (4.4), we can take the same approach as in Section 3.2. Again optimal policies for discounted risk sensitive control problems are nonstationary, so we take logarithms in (4.4) and use the duality [10], via the Legendre transformation, between the log moment generating function and the relative entropy, to convert (4.4) into the following optimality equation for an average cost game:

$$\lambda + W(\alpha) = \min_{u \in U} \sup_{\xi \in P(Y)} \sum_{j=0}^{1} \xi^j \{ -log(2\xi^j) + log|M(j,u) \begin{bmatrix} 1-\alpha \\ \alpha \end{bmatrix} | + W(f(\alpha,j,u)) \}.$$

Approximation with the corresponding discounted equations, as in Section 3.2, yields the existence of solutions to (4.4).

Proposition 4.9 [24]. There exist a number λ and a bounded function $W : [0,1] \to \mathbb{R}$ such that (4.4) is satisfied.

Structural Results.

Using arguments similar to those in [13], we can obtain results on the structure of the optimal policy.

Lemma 4.10 [24]. If every average cost optimal policy replaces at $\alpha \in [0,1]$, then every average cost optimal policy replaces in the interval $[\alpha, 1]$.

Lemma 4.11 [24]. It is average cost optimal to produce in the interval $[0, f(0,0,0)]$.

Condition C.1. For each $\beta \in (0,1)$,

$$\frac{1}{\gamma} C \le (1-\beta)[\frac{1}{\gamma} R + \beta W_\beta(0)].$$

Theorem 4.12 [24]. a) If (C.1) holds, then it is optimal to produce for all $\alpha \in [0,1]$.

b) If (C.1) does not hold, then there exists $\alpha_\gamma \in [0,1)$ such that it is optimal to produce in $[0,\alpha_\gamma)$ and repair in $[\alpha_\gamma,1]$.

Thus a simple threshold or "bang-bang" is optimal. Simulations are presented in [24] to compare the optimal policies for the risk sensitive and risk neutral cost criteria, and to study how the policies vary as a function of the risk factor γ^{-1}.

5 Alternative Risk Sensitive Approach

The risk sensitive approaches discussed above are generalizations of the risk neutral approach, in the sense that they seek to minimize a cost functional that is a generalization of the risk neutral cost functional. This works well for finite horizon and average cost risk sensitive problems, and results have been presented that correspond to those in the risk neutral case. However, as noted in Section 3.2, the minimization of the risk sensitive discounted cost functional (3.3) results in nonstationary optimal policies [8].

Instead of generalizing the expression for the cost to be minimized, one can alternatively generalize the risk-neutral dynamic programming equations characterizing the value function. This leads to a formulation for which, on the infinite horizon and with discounting and stationary costs, there exists a stationary optimal policy (see [8], [9], [21], [28]). This alternative formulation does *not* involve the optimization of a single cost (or utility) function for the entire path of the process; indeed, there is considerable debate in the decision theory and economics literature about whether the optimization of a single expected utility function is "rational" (see [28] and the references therein). In this alternative formulation, the control at each time is chosen to minimize an immediate cost, plus the discounted "certain equivalent" return from future stages, resulting in the following dynamic programming equation for the value function h^γ in the completely observed, finite state case:

$$h^\gamma(x) = \min_{u\in U} c(x,u) + \beta\gamma \log \sum_{z\in X} P_{xz}(u)\exp(\frac{1}{\gamma}h^\gamma(z)) \tag{5.1}$$

Existence of stationary optimal policies, as well as policy iteration algorithms, are shown in [8]. Extensions of the theory and the partially observed case are discussed in [9].

The discounted dynamic programming equation (5.1) is similar to (3.6) for the discounted game problem of Section 3.2. Indeed, Eagle [11] studied (3.6) in a context quite similar to that discussed in this section. An intriguing question is that of developing a more fundamental understanding of the relationship between the discounted problems discussed in this section and the discounted games discussed in Section 3.2.

References

[**1**] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, Discrete-time controlled Markov processes with average cost criterion: a survey, *SIAM J. Control and Optim.* (1993) 31, 282-344.

[**2**] J. S. Baras and M. R. James, Robust and risk-sensitive output feedback control for finite state machines and hidden Markov models, *J. Math. Systems, Estimation and Control* (to appear).

[**3**] A. Bensoussan and J. H. Van Schuppen, Optimal control of partially observable stochastic systems with exponential of integral performance index, *SIAM J. Control and Optim.* (1985) 23, 599-613.

[**4**] D.P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models,* Prentice-Hall, Englewood Cliffs, 1987.

[**5**] V. S. Borkar, On minimum cost per unit of time control of Markov chains, *SIAM J. Cont. and Optim.* 22 (1984), 965-978.

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

[**7**] R. Cavazos-Cadena and L. I. Sennott, Comparing recent assumptions for the existence of average optimal stationary policies, *Oper. Res. Lett.* 11 (1992), 33-37.

[**8**] K.-J. Chung and M. J. Sobel, Discounted MDP's: Distribution functions and exponential utility maximization, *SIAM J. Control and Optim.* (1987) 25, 49-62.

[**9**] S. Coraluppi and S. I. Marcus, Risk-sensitive control of Markov decision processes, *Proc. 1996 Conf. on Information Science and Systems*, 934-939.

[**10**] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, forthcoming, to be published by John Wiley & Sons.

[**11**] J. N. Eagle II, A utility criterion for the Markov decision process, Ph. D. thesis, Stanford University, Stanford, CA, 1975.

[**12**] C.-H. Fan, J. L. Speyer and C. R. Jaensch, Centralized and decentralized solutions fo the linear-exponential-Gaussian problem, *IEEE Transactions on Automatic Control* (1994) 39, 1986-2003.

[13] E. Fernández-Gaucherand, A. Arapostathis, and S.I. Marcus, On the average cost optimality equation and the structure of optimal Policies for partially observable Markov decision processes, *Annals of Operations Research* (1991) 29, 439–470.

[14] E. Fernández-Gaucherand, A. Arapostathis and S.I. Marcus, Analysis of an adaptive control scheme for a partially observed controlled Markov chain, *IEEE Transactions on Automatic Control* (1993) 38, 987-993.

[15] E. Fernández-Gaucherand and S. I. Marcus, Risk-sensitive optimal control of hidden Markov models: structural results, *IEEE Transactions on Automatic Control* (to appear).

[16] W. H. Fleming and D. Hernández-Hernández, Risk sensitive control of finite state machines on an infinite horizon I, *SIAM J. Control and Optim.* (to appear).

[17] W. H. Fleming and D. Hernández-Hernández, Risk sensitive control of finite state machines on an infinite horizon II, Technical Report, Division of Applied Mathematics, Brown University.

[18] W. H. Fleming and W. M. McEneaney, Risk-sensitive control and differential games, *Springer Lecture Notes in Control and Info. Sci.* No. 184, 1992, 185-197.

[19] W. H. Fleming and W. M. McEneaney, Risk-sensitive control on an infinite horizon, *SIAM J. Control and Optim.* (1995) 33,1881-1915.

[20] K. Glover and J. C. Doyle, State-space formulae for all stabilizing controllers that satisfy an H_∞-norm bound and relations to risk sensitivity, *Systems and Control Lett.* (1988) 11, 167-172.

[21] L. P. Hansen and T. J. Sargent, discounted linear exponential quadratic Gaussian control, *IEEE Transactions on Automatic Control* (1995) 40, 968-971.

[22] D. Hernández-Hernández and S. I. Marcus, Risk-sensitive control of Markov processes in countable state space, *Systems and Control Lett.* (to appear).

[23] D. Hernández-Hernández and S. I. Marcus, Existence of risk sensitive optimal stationary policies for controlled Markov processes, Technical Report, Institute for Systems Research, University of Maryland.

[24] D. Hernández-Hernández, S. I. Marcus, and P. Fard, Analysis of a risk sensitive control problem for hidden Markov chains, Technical Report, Institute for Systems Research, University of Maryland.

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

[26] D. H. Jacobson, Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games, *IEEE Transactions on Automatic Control* (1973) 18, 124-131.

[27] M. R. James, J. S. Baras and R. J. Elliott, Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems, *IEEE Transactions on Automatic Control* (1994) 39, 780-792.

[28] D. M. Kreps and E. L. Porteus, Temporal resolution of uncertainty and dynamic choice theory, *Econometrica* (1978) 46, 185-200.

[29] P.R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, 1986.

[30] W.S. Lovejoy, On the convexity of policy regions in partially observed systems, *Operations Research* (1987) 35, 619-621.

[31] T. Runolfsson, The equivalence between infinite horizon control of stochastic systems with exponential-of-integral performance index and stochastic differential games, *IEEE Transactions on Automatic Control* (1994) 39, 1551-1563.

[32] R.D Smallwood and E.J. Sondik, The optimal control of partially observable Markov processes over a finite horizon, *Operations Research* (1973) 21, 1071-1088.

[33] C.C. White, A Markov quality control process subject to partial observation, *Management Science* (1977) 23, 843-852.

[34] P. Whittle, *Risk-Sensitive Optimal Control*, John Wiley & Sons, New York, 1990.

View publication stats