Risk-Sensitive and Minimax Control of Discrete-Time, Finite-State Markov Decision Processes*

Stefano P. Coraluppi[1] and Steven I. Marcus[2]

Abstract

This paper analyzes a connection between risk-sensitive and minimax criteria for discrete-time, finite-states Markov Decision Processes (MDPs). We synthesize optimal policies with respect to both criteria, both for finite horizon and discounted infinite horizon problem. A generalized decision-making framework is introduced, which includes as special cases a number of approaches that have been considered in the literature. The framework allows for discounted risk-sensitive and minimax formulations leading to stationary optimal policies on the infinite horizon. We illustrate our results with a simple machine replacement problem.

Key Words: Stochastic Control, Risk-Sensitive Control, Minimax Control, Markov Decision Processes

1 Introduction

In the classical, risk-neutral approach to stochastic control, one seeks to minimize the expected total cost (or average cost) incurred in the evolution of a dynamical system. Risk-sensitive control is a generalization of this approach whereby we consider higher order moments of the probability distribution for the total cost as well. In minimax control, one is interested in minimizing the worst-case behavior of a dynamical system.

An early formulation of the risk-sensitive control problem is due to [HM72]. In the LQG setting, the problem was first studied by [Jac73], where it was found that in the risk-sensitive setting, the certainty equivalence principle does not hold in its original form. Extensions to the partially observed setting include [Whi81] and [BS85]. A

*Corresponding author Steven I. Marcus. Tel. 301-405-7589; Fax 301-314-9920; E-mail marcus@isr.umd.edu.

[1]Department of Electrical Engineering and Institute for Systems Research, University of Maryland at College Park, MD 20742, U.S.A.

[2]Department of Electrical Engineering and Institute for Systems Research, University of Maryland at College Park, MD 20742, U.S.A.

somewhat surprising result is that the conditional distribution of the state given past observations does not constitute an information state.

A good survey of work in nonlinear risk-sensitive control is given by [McE96a] and [McE96b]. The partially-observed MDP setting has been studied in [BJam], where an information state and dynamic programming equations for the value function on the finite horizon are introduced. Structural results for the value function are due to [FGMar].

Early work in minimax control of stochastic systems includes [BR71], where the connection between stochastic and deterministic descriptions of uncertainty is addressed. In the LQG setting, a connection between risk-sensitive control and H_∞ control is established in [GD88]. The connection between minimax and robust control is explored in [BB95]. In [BJam], a finite-state robust control problem is studied as the small-noise limit of a particular risk-sensitive control problem.

An interesting fact both in risk-sensitive and minimax control is that in general, on the infinite horizon and with stationary costs, there does not exist a stationary optimal policy. This is the case in the finite-state MDP setting as well. Dynamic programming equations in the full state observations case are derived in [CS87]. Alternate approaches to risk-sensitive control which lead to stationary optimal policies are developed in [Por75], [KP78], and [Eag75]. An alternate approach in the LQG setting is developed in [HS95]. Average cost approaches, which also lead to optimal stationary policies on the infinite horizon, are pursued in [FHH(1)], [FHH(2)], [HHM96], [HHM97].

In this paper, we analyze the large-risk-limit connection between the risk-sensitive and the minimax control problems in the MDP setting. The minimax control problem can be addressed by exploiting this connection. We synthesize optimal risk-sensitive and minimax policies on the finite horizon, and derive dynamic programming equation on the infinite horizon with discounted costs. A sufficiently large finite horizon approximation to the infinite horizon problem can be used to obtain near-optimal policies both for risk-sensitive and minimax criteria.

Further, we introduce a generalized decision-making framework which includes as special cases a number of approaches that have been considered in the literature, and extend these approaches to the minimax setting. We illustrate our results with a machine replacement problem that has been used as a benchmark example in the literature (see [FGMar]).

2 Risk-Sensitive and Minimax Control

We consider the class of discrete-time MDPs with finite state space X, finite control space U, and finite observation space Y. We denote the cardinality of these spaces by $|X|$, $|U|$, and $|Y|$. The probability transition matrix $P(u)$ is defined by $P_{ij}(u) = pr(x_{k+1} = j | x_k = i, u_k = u)$, and the observation matrix $Q(u)$ is defined by $Q_{ij}(u) = pr(y_k = j | x_k = i, u_{k-1} = u)$. We define $c_k(x_k, u_k) \geq 0$ to be the (possibly discounted) cost incurred by the system at time $k \geq 0$, given that it is in state $x_k \in X$ and that control $u_k \in U$ is used. If there is a finite horizon size N, there is a terminal cost

$c_N(x_N) \geq 0$. A partial sum of costs is denoted by $C_{i,N} = \sum_{k=i}^{k=N-1} c_k(x_k, u_k) + c_N(x_N)$. The vector of terminal costs is denoted by c_N.

The *risk-neutral* objective is given by

$$J(\mu, \pi_0) = E^{\mu,\pi_0}[\sum_k c_k(x, u)], \tag{1}$$

where μ is a non-anticipative policy and π_0 is the probability distribution on the states of the system at time $k = 0$. The *risk-sensitive* objective is given by

$$J^\gamma(\mu, \pi_0) = \frac{1}{\gamma} \log E^{\mu,\pi_0}[\exp{(\gamma \sum_k c_k(x, u))}]. \tag{2}$$

For small γ, (2) takes the form

$$J^\gamma(\mu, \pi_0) \simeq E^{\mu,\pi_0}[\sum_k c_k(x, u)] + \frac{\gamma}{2} Var^{\mu,\pi_0}[\sum_k c_k(x, u)], \tag{3}$$

and in the limit $\gamma \to 0$, (2) reverts to the risk-neutral objective (1). The parameter γ allows one to incorporate an aversion or preference for risk, or variability in the cost incurred in the system's evolution. For $\gamma > 0$, we are penalized for variability in the cost incurred, so we say that we have a *risk-averse* objective.

An equivalently objective to (2) is given by

$$\hat{J}^\gamma(\mu, \pi_0) = E^{\mu,\pi_0}[\exp{(\gamma \sum_k c_k(x, u))}]. \tag{4}$$

In [BJam], an information state process for the MDP with respect to criterion (4) is defined, satisfying the following recursion:

$$\begin{aligned} \sigma_0^\gamma &= \pi_0, & (5) \\ \sigma_{k+1}^\gamma &= |Y| \sigma_k^\gamma D^\gamma(k, u_k) \bar{Q}(y_{k+1}, u_k), & (6) \end{aligned}$$

where

$$D_{ij}^\gamma(k, u) := P_{ij}(u) \exp{(\gamma c_k(i, u))}, \tag{7}$$

and $\bar{Q}(\cdot, \cdot)$ is a diagonal matrix with $\bar{Q}_{ii}(y, u_k) = pr(y_{k+1} = y | x_{k+1} = i, u_k = u)$. The information state belongs to the space $R_+^{|X|}$, where R_+ is the space of non-negative real numbers. On the finite horizon, the value function associated with this information state is given by

$$S_{k,N}^\gamma(\sigma) := \inf_{\mu \in M} E^\dagger[\sigma_N^\gamma \cdot \exp(\gamma c_N) | \sigma_k^\gamma = \sigma]. \tag{8}$$

where the *exp* operator is defined component-wise, M denotes the set of non-anticipative policies, and $\dagger$ denotes a reference probability measure, under which all observations $y \in Y$ are independent and equiprobable at every time k. Dynamic programming equations for (8) are given by

$$\begin{aligned} S_{N,N}^\gamma(\sigma^\gamma) &= \sigma^\gamma \cdot \exp(\gamma c_N), & (9) \\ S_{k,N}^\gamma(\sigma^\gamma) &= \min_{u \in U} E^\dagger[S_{k+1,N}^\gamma(p \sigma^\gamma D^\gamma(k, u) \bar{Q}(y_{k+1}, u)]. & (10) \end{aligned}$$

It has been shown in [FGMar] that $S^{\gamma}_{k,N}(\cdot)$ is a concave and piecewise-linear function. These structural properties together with a normalized information state can be exploited to develop an algorithm to synthesize an optimal policy, similar to the algorithm given in [SS73] for risk-neutral control (with a minor correction in [Lov89]). See [Cor97] for details.

The *minimax* objective is given by

$$\bar{J}(\mu, \pi_0) = \sup_{\omega \in \Omega^{\mu}} \sum_k c_k(x_k, u_k), \tag{11}$$

where Ω^{μ} is the set of trajectories of the form $(x_0, u_0, x_1, u_1, \ldots)$ that occur with non-zero probability under policy μ. Note that, with respect to the minimax objective, the probability with which each trajectory occurs under a fixed policy μ is significant only to the extent that it is zero or non-zero.

The following result will be useful in establishing a connection between the risk-sensitive and minimax criteria. Its proof is similar to that of the Varadhan-Laplace Lemma, given e.g. in [BJam].

Lemma 1 (Modified Varadhan-Laplace Lemma). Let F^{γ}, F be real valued functions defined on a finite set Ω, where $\forall \omega \in \Omega,\ F(\omega) = \lim_{\gamma \to \infty} F^{\gamma}(\omega)$. Also, let $p(\omega)$ be a nonnegative real number $\forall \omega \in \Omega$, independent of γ. Then

$$\lim_{\gamma \to \infty} \frac{1}{\gamma} \log \sum_{\omega \in \Omega} p(\omega) \exp\left[\gamma F^{\gamma}(\omega)\right] = \max_{\omega \in \Omega, p(\omega) \neq 0} F(\omega). \tag{12}$$

Using Lemma 1, it can be shown that, on the finite horizon, $\lim_{\gamma \to \infty} J^{\gamma}(\cdot,\cdot) = \bar{J}(\cdot,\cdot)$. That is, the large-risk limit of the risk-sensitive objective is the minimax objective. Let us define a statistic for the MDP by

$$s_k := \lim_{\gamma \to \infty} \frac{1}{\gamma} \log \sigma^{\gamma}_k, \ \forall k, \tag{13}$$

where the *log* operator is defined component-wise. Again using Lemma 1, it can be shown that the statistic satisfies the following recursion, where by $s[x]$ we mean the xth component of vector s:

$$s_0[x] = \begin{cases} 0 & \text{if } \pi_0[x] \neq 0, \\ -\infty & \text{otherwise,} \end{cases} \tag{14}$$

$$s_{k+1}[x'] = f(s_k, u_k, y_{k+1}). \tag{15}$$

The function $f(\cdot,\cdot,\cdot)$ is given by

$$f(s_k, u_k, y_{k+1}) = \begin{cases} \max_{x \in \tilde{X}(x', u_k)} [s_k[x] + c_k(x, u_k)] & \text{if } \tilde{X}(x', u_k) \neq \emptyset,\ x' \in \tilde{Y}(y_{k+1}, u_k) \\ -\infty & \text{otherwise,} \end{cases} \tag{16}$$

where $\tilde{X}(x', u_k)$ is the set of states at time k from which, using control $u_k \in U$, there is a nonzero probability that the state of the system at time $k+1$ will be x'; $\tilde{Y}(y_{k+1}, u_k)$ is the set of states at time $k+1$ that can result in observation y_{k+1} at time $k+1$, if the control at time k is u_k.

It can be shown that the statistic and the objective (11) on the finite horizon are related by the following:

$$\bar{J}(\mu, \pi_0) = \max_{y_1, \ldots, y_N} \max_{i \in X} s_N[i]. \tag{17}$$

This motivates the following definition for the value function:

$$W_{k,N}(s) := \min_{\mu \in M} \max_{y_{k+1}, \ldots, y_N} \max_{i \in X} s_N[i], \text{ where } s_k = s. \tag{18}$$

Indeed, we have

$$W_{0,N}(s_0) = W_{0,N}(\lim_{\gamma \to \infty} \frac{1}{\gamma} \log \pi_0) = \min_{\mu \in M} \bar{J}(\mu, \pi_0). \tag{19}$$

The value function at time k can be thought of as the worst case total cost incurred in the system's evolution, given an information state at time k, and given that an optimal policy is used thereafter.

The following result establishes that the statistic satisfying (14), (15) is an information state, and that there exists an optimal separated policy that can be computed by using the dynamic programming equations for the value function (18). First, we introduce the following notation for the set of all information states. Define $\tilde{R}_+^{|X|} := \{R_+, -\infty\}^{|X|}$.

Theorem 1 (Minimax Finite Horizon Dynamic Programming). The value function satisfies the following, $\forall s \in \tilde{R}_+^{|X|}$:

$$\begin{aligned} W_{N,N}(s) &= \max_{i \in X} s[i], & (20)\\ W_{k,N}(s) &= \min_{u \in U} \max_{y \in Y} W_{k+1,N}(f(s, u, y)), & (21) \end{aligned}$$

A policy that achieves the minimum in equations (20) and (21), also achieves the minimum in (18). Furthermore, the policy is separated and is optimal with respect to (11).

Proof. Equation (20) follows immediately from (18). For $k < N$, using Lemma 1 and (13), we have

$$\begin{aligned} W_{k,N}(s) &= \min_{\mu \in M} \lim_{\gamma \to \infty} \frac{1}{\gamma} \log E^{\mu}[\exp(\gamma C_{k,N}) | \sigma_k^{\gamma} = \exp(\gamma s)] \\ &= \lim_{\gamma \to \infty} \frac{1}{\gamma} \log \min_{\mu \in M} E^{\mu}[\exp(\gamma C_{k,N}) | \sigma_k^{\gamma} = \exp(\gamma s)] \\ &= \lim_{\gamma \to \infty} \frac{1}{\gamma} \log S_{k,N}^{\gamma}(\exp(\gamma s)). & (22) \end{aligned}$$

Thus we have

$$\begin{aligned} W_{k,N}(s) &= \lim_{\gamma\to\infty} \frac{1}{\gamma} \log \min_{u\in U} E^{\dagger}[S^{\gamma}_{k+1,N}(|Y| \exp(\gamma s) D^{\gamma}(k,u)\bar{Q}(y,u))] \\ &= \min_{u\in U} \lim_{\gamma\to\infty} \frac{1}{\gamma} \log E^{\dagger}[S^{\gamma}_{k+1,N}(|Y| \exp(\gamma s) D^{\gamma}(k,u)\bar{Q}(y,u))] \\ &= \min_{u\in U} \lim_{\gamma\to\infty} \frac{1}{\gamma} \log \sum_{y\in Y} |Y| S_{k+1,N}(\exp(\gamma s^{\gamma}_{k+1})), \end{aligned} \tag{23}$$

where s^{γ}_{k+1} is such that $\exp(\gamma s^{\gamma}_{k+1}) = |Y| \exp(\gamma s) D^{\gamma}(k,u)\bar{Q}(y,u)$. Note that $\lim_{\gamma\to\infty} s^{\gamma}_{k+1} = s_{k+1}$, as given by equation (15). Finally, using Lemma 1 and (22), we conclude

$$W_{k,N}(s) = \min_{u\in U} \max_{y\in Y} W_{k+1,N}(f(s,u,y)). \tag{24}$$

It remains to show the optimality of a separated policy. Note that the minimization in (21) depends on past observations only through the information state. Thus, setting $k = 0$ we see that a total cost of $W_{0,N}(s_0)$ is achieved by using a separated policy. It follows from (19) that the policy is optimal in the larger class M of all admissible policies. □

In risk-neutral and risk-sensitive control, the determination of optimal policies for partially observed MDPs typically involves the use of structural results for the value function. Without such results, the minimization in (10) over a continuum of information states (the unit simplex), is intractable. In the minimax control setting, the situation is greatly simplified since, on the finite horizon, we need only consider a finite number of information states. At time $k = 0$, there are $2^{|X|} - 1$ values that the information state s_0 can take, corresponding to all possible subsets of X of feasible initial states. At time $k > 0$, in the worst case there are $(2^{|X|} - 1)(|U| \cdot |Y|)^k$ feasible information states. A possible scheme for determining optimal policies on the finite horizon is the following:

1. Generate all information states of interest.
2. Use the dynamic programming equations (20), (21) to find the optimal control at each state of interest.

The use of this scheme will be illustrated in Section 5.

3 The Infinite Horizon

One way to insure that the objectives (1), (2), and (11) are bounded on the infinite horizon by introducing a discounted cost structure. That is, we set $c_k(\cdot,\cdot) = \beta^k c(\cdot,\cdot)$, where $0 < \beta < 1$. In [CS87] it is shown that the limit

$$\hat{S}^{\gamma}_{k}(x) := \lim_{N\to\infty} \hat{S}^{\gamma}_{k,N}(x) \tag{25}$$

exists, for all $x \in X$ and $\gamma > 0$, where $\hat{S}^{\gamma}_{k,N}(x),\ x \in X$ is the value function in the case of full state observations. Furthermore, the infinite horizon value function can be characterized as follows:

$$\hat{S}^{\gamma}_0 = \min_{u \in U} \{D^{\gamma}(0,u)\hat{S}^{\beta\gamma}_0\}, \tag{26}$$

where the minimum is taken separately for each component of the vector equation. Analogously, in the partially observed setting we have the following.

Theorem 2 (Risk-Sensitive Infinite Horizon Dynamic Programming). For all $\sigma \in \Re^{|X|}_{+}$, and $\gamma > 0$, define

$$S^{\gamma}_k(\sigma) := \lim_{N\to\infty} S^{\gamma}_{k,N}(\sigma), \tag{27}$$

where $S^{\gamma}_{k,N}$ is defined in (8). The limit in (27) exists, and

$$S^{\gamma}_0(\sigma) = \min_{u \in U} E^{\dagger}[S^{\beta\gamma}_0(p\sigma D^{\gamma}(0,u)\bar{Q}(y_1,u))]. \tag{28}$$

Proof. The existence of the limit in (27) can be established as in ([CS87]). Now, for finite N, we have

$$S^{\gamma}_{0,N}(\sigma) = \min_{u \in U} E^{\dagger}[S^{\gamma}_{1,N}(p\sigma D^{\gamma}(0,u)\bar{Q}(y_1,u))]$$

from (10). Letting $N \to \infty$, we have

$$\begin{aligned} S^{\gamma}_0(\sigma) &= \lim_{N\to\infty}\min_{u\in U} E^{\dagger}[S^{\gamma}_{1,N}(p\sigma D^{\gamma}(0,u)\bar{Q}(y_1,u))] \\ &= \min_{u\in U} E^{\dagger}[\lim_{N\to\infty} S^{\gamma}_{1,N}(p\sigma D^{\gamma}(0,u)\bar{Q}(y_1,u))] \\ &= \min_{u\in U} E^{\dagger}[S^{\gamma}_{1}(p\sigma D^{\gamma}(0,u)\bar{Q}(y_1,u))] \\ &= \min_{u\in U} E^{\dagger}[S^{\beta\gamma}_{0}(p\sigma D^{\gamma}(0,u)\bar{Q}(y_1,u))], \end{aligned} \tag{29}$$

using continuity of $S^{\gamma}_{k,N}(\cdot)$ and the finiteness of the output space Y. □

Proceeding in a similar fashion for the minimax objective, we introduce the following infinite horizon value function:

$$W_k(s) := \lim_{N\to\infty} W_{k,N}(s). \tag{30}$$

We can verify that the limit in (30) is well-defined by recalling that $W_{k,N} = \lim_{\gamma\to\infty}\frac{1}{\gamma}\log S^{\gamma}_{k,N}(\exp(\gamma s))$, and $\lim_{N\to\infty} S_{k,N}$ is well-defined. Thus

$$W_k(s) = \lim_{\gamma\to\infty}\frac{1}{\gamma}\log S^{\gamma}_k(\exp(\gamma s)) \tag{31}$$

We can relate the value function to the criterion (11) by taking the limit in (19) as $N \to \infty$. We obtain:

$$W_0(s_0) = \inf_{\mu \in M} \bar{J}(\mu, \pi_0) \tag{32}$$

The following result characterizes the infinite horizon value function.

Theorem 3 (Minimax Infinite Horizon Dynamic Programming). The value function (30) satisfies the following, $\forall s \in \tilde{R}_+^{|X|}$:

$$W_0(s) = \min_{u \in U} \max_{y \in Y} \beta W_0(\frac{f(s,u,y)}{\beta}). \tag{33}$$

Proof. First, we derive a relationship between time-shifted value functions analogous to (29) in the risk-sensitive setting.

$$\begin{aligned} W_1(s) &= \lim_{\gamma \to \infty} \frac{1}{\gamma} \log S_1^{\gamma}(\sigma^{\gamma}) \\ &= \lim_{\gamma \to \infty} \frac{1}{\gamma} \log S_0^{\beta\gamma}(\sigma^{\gamma}) \\ &= \lim_{\gamma \to \infty} \beta \frac{1}{\beta\gamma} \log S_0^{\beta\gamma}(\sigma^{\gamma}) \\ &= \beta W_0(s'), \end{aligned}$$

where $s' = \lim_{\gamma \to \infty} \frac{1}{\beta\gamma} \log(\sigma^{\gamma})$. It follows that

$$W_1(s) = \beta W_0(\frac{s}{\beta}). \tag{34}$$

We have

$$\begin{aligned} W_0(s) &= \min_{u \in U} \max_{y \in Y} W_1(f(s,u,y)) \\ &= \min_{u \in U} \max_{y \in Y} \beta W_0(\frac{f(s,u,y)}{\beta}), \end{aligned}$$

using (34). □

In risk-neutral control, a stationary optimal policy can be determined through policy or value iteration techniques. Unfortunately, both in the risk-sensitive and the minimax settings in general there does not exist a stationary optimal policy. Thus, the optimal policies satisfying equations (28) and (33) are difficult to determine. Given a tolerance bound $\epsilon > 0$, we can consider the truncation of the infinite horizon to a finite horizon of $N = \max\{\lceil \xi \rceil, 1\}$, where $\xi = \frac{\log[\frac{(1-\beta)\epsilon}{\|c\|}]}{\log \beta}$, and $\| c \| := \max_{x \in X, u \in U} |c(x,u)|$. Both for risk-sensitive and minimax criteria, if we solve the finite horizon dynamic programming equations with horizon size N and no terminal cost, and use a fixed, arbitrary policy thereafter, the resulting objectives (2) and (11) are within ϵ of optimal. See ([Cor97]) for details.

4 A Generalized Decision-Making Framework

Motivated by the the lack of stationary optimal policies for discounted risk-sensitive and minimax criteria, and the complexity associated with solving the dynamic programming equations (9), (10) or (20), (21) for a large horizon N, we would like to formulate optimal risk-sensitive and minimax decision-making in a more general setting, leading to stationary discounted optimal policies on the infinite horizon. An additional motivation is provided by decision theorists, many of whom argue (see e.g. [EZ89]) that a normative theory for decision-making must lead to stationary optimal policies on the infinite horizon.

Assume that the state of the MDP is observed. On the finite horizon, the value function corresponding to the risk-sensitive criterion (2) can be defined as

$$s^{\gamma}_{k,N}(i) := \min_{\mu} \frac{1}{\gamma} \log E^{\mu}[\exp(\gamma C_{k,N})|x_k = i], \; i \in X \tag{35}$$

Recall that $C_{k,N} = \sum_{j=k}^{N-1} c_j(x_j, u_j) + c_N(x_N)$. The dynamic programming equations for (35) are given by

$$s^{\gamma}_{k,N}(i) = \min_{u \in U}\{c_k(i,u) + \frac{1}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma s^{\gamma}_{k+1,N}(j))]\}, \tag{36}$$

$$s^{\gamma}_{N,N}(i) = c_N(i). \tag{37}$$

In the small-risk limit, $\gamma \to 0$, (36), (37) revert to the usual risk-neutral dynamic programming equations. On the infinite horizon, we have

$$s^{\gamma}_{k}(i) = \min_{u \in U}\{c_k(i,u) + \frac{1}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma s^{\gamma}_{k+1}(j))]\}, \; k = 0, \ldots. \tag{38}$$

If $c_k(\cdot,\cdot) = \beta^k c(\cdot,\cdot)$, it can be shown that time-shifted value functions are related as follows:

$$s^{\gamma}_{k+1}(\cdot) = \beta s^{\beta\gamma}_{k}(\cdot). \tag{39}$$

Equation (39) also reverts to a well-known relationship in the risk-neutral case:

$$s^{0}_{k+1}(\cdot) = \beta s^{0}_{k}(\cdot). \tag{40}$$

A more general set of optimality equations than (36), (37) can be defined as follows:

$$h^{\gamma}_{k,N}(i) = \min_{u \in U}\{c_k(i,u) + \frac{\beta'}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma\beta'' h^{\gamma}_{k+1,N}(j))]\}, \tag{41}$$

$$h^{\gamma}_{N,N}(i) = c_N(i). \tag{42}$$

An interpretation for these optimality equations is that the value function at time k equals the cost incurred at time k, plus a (possibly discounted) contribution accounting for future costs. Note that if we set $\beta' = \beta'' = 1$, we revert to the classical risk-sensitive dynamic programing equations. If we set $\beta = \beta'' = 1$, we obtain the formulation that has been studies in a series of papers including [Por75] and [KP78], which we refer to as the *Porteus* formulation. A similar formulation in the LQG setting has been proposed recently in [HS95]. If we set $\beta = \beta' = 1$, we obtain the formulation introduced in [Eag75], which we refer to as the *Eagle* formulation.

On the infinite horizon, setting $c_k(\cdot,\cdot) = \beta^k c(\cdot,\cdot)$, the generalized optimality equation is given by

$$h_k^\gamma(i) = \min_{u\in U}\{\beta^k c(i,u) + \frac{\beta'}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma\beta'' h_{k+1}^\gamma(j))]\},\ k = 0,\ldots. \tag{43}$$

Once again we obtain the classical, Porteus, and Eagle formulations as special cases of (43). A key feature of the generalized formulation (43) is that it is sufficient for one of β, β', and β'' to be less than 1, provided the others are set to 1, to insure boundedness of the value function h_k^γ. Thus, by setting either β' or β'' to be less than one, we can set $\beta = 1$. It can then be shown that $h_k^\gamma(\cdot) = h^\gamma(\cdot)$, that is we have a time-invariant value function, and furthermore there is a stationary policy that achieves the minimum in (15). It can further be shown that policy and value iteration techniques can be used to synthesize an optimal policy. See [Cor97] for details, and for extensions to the partial state observations setting.

The nature of the discount factors β, β', and β'' can be better understood by considering the small-risk limit, $\gamma \to 0$, of (43). We obtain the following:

$$h_k^0(i) = \min_{u\in U}\{\beta^k c(i,u) + \beta'\beta''\sum_j P_{ij}(u)h_{k+1}^0(j)\},\ k = 0,\ldots. \tag{44}$$

Note that this optimality equation is more general than the risk-neutral dynamic programming equation. On the other hand, each of the three special cases of (43) that we have considered (classical, Porteus, Eagle) is equivalent to risk-neutral control in the small-risk limit.

A generalized minimax formulation is given by

$$\bar{h}_{k,N}(i) = \min_{u\in U}\{c_k(i,u) + \beta'\beta''\max_{j\in\tilde{X}'(i,u)}\bar{h}_{k+1,N}(j)\}, \tag{45}$$

$$\bar{h}_{N,N}(i) = c_N(i), \tag{46}$$

where once again $\tilde{X}'(i,u)$ is the set of states that the system reaches in one transition with nonzero probability, given that it is in state i and control u is used. On the infinite horizon and with $c_k(\cdot,\cdot) = \beta^k c(\cdot,\cdot)$, the generalized minimax formulation is given by

$$\bar{h}_k(i) = \min_{u\in U}\{\beta^k c(i,u) + \beta'\beta''\max_{j\in\tilde{X}'(i,u)}\bar{h}_{k+1}(j)\}. \tag{47}$$

It can be shown that the generalized minimax formulation is the large-risk limit of the generalized risk-sensitive formulation. It follows that when $\beta = 1$ and at least one of β', β'' is less than 1, once again the value function is time-invariant, and there exists a stationary optimal policy that can be determined by policy or value iteration techniques.

An interesting consequence of introducing the additional discount parameters β' and β'' in the risk-sensitive formulation is that, unlike (36), (37), the equations (41), (42) are not dynamic programming equations. By this we mean that, in general, a policy $\mu^\star$ achieving the minimum on the *r.h.s.* of equations (41), (42) does not minimize a criterion of expected utility form. More precisely, in general there does not exist a $U : \Re_+ \to \Re_+$, such that the objective $E^\mu[U(\sum_k c_k(x_k, u_k))]$ is minimized by policy $\mu^\star$. The same comment applies to the infinite horizon optimality equation (43). This can be understood in light of the axiomatic foundation of Utility Theory (see e.g. [HS84]), and some dynamic extensions discussed in [KP78].

5 Machine Replacement Example

Let us consider the following benchmark problem which has appeared in the literature (see [FGMar]). We have state space $X = \{0, 1\}$, observation space $Y = \{0, 1\}$, and control space $U = \{0, 1\}$. The probability transition matrix and output matrix are given by

$$P(0) = \begin{bmatrix} 1-\theta & \theta \\ 0 & 1 \end{bmatrix}, \quad P(1) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad 0 < \theta < 1, \tag{48}$$

and

$$Q = \begin{bmatrix} q & 1-q \\ 1-q & q \end{bmatrix}. \tag{49}$$

The MDP models an error prone manufacturing or communication system. The working state is $x = 0$, and the failed state is $x = 1$. The control options are to keep ($u = 0$) or repair ($u = 1$). The cost incurred in the system's evolution is defined by $c(0, 0) = 0$, $c(1, 0) = C$, and $c(x, 1) = R$, $x \in X$. The cost to repair exceeds the cost associated with operating the faulty unit, that is $R > C$.

The probability transition matrices can be understood as follows. If the system is working and we do not replace it, there is a probability θ that it will be broken at the next time unit. A broken unit will stay broken if it is not replaced. If the system is replaced, it is certain to be in the working state at the next time unit. The quality of observation is given by $q > 0$.

5.1 Finite Horizon

Let us assume that the system evolves on the finite horizon, with $N = 3$, and that there is no terminal cost. Let us also assume that $q < 1$, that is we do not have perfect state observations.

We now implement the methodology introduced in Section 2 to determine a separated optimal policy. We consider all possible initial distributions on the state of the MDP. These can be divided into three classes, each leading to a unique initial information state s_0:

1. $\pi_0 = [\pi_0[0], \pi_0[1]], \ 0 < \pi_0[0], \ \pi_0[1] < 1 \Rightarrow s_0 = [0, 0]$.

$k = 0$	$s_0 \in \{[0, 0]\}$
$k = 1$	$s_1 \in \{[0, C], [R, -\infty]\}$
$k = 2$	$s_2 \in \{[0, 2C], [R + C, -\infty], [R, R], [2R, -\infty]\}$
$k = 3$	$s_3 \in \{[0, 3C], [2C + R, -\infty], [R + C, R + C], [2R + C, -\infty],$ $[R, R + C], [2R, -\infty], [2R, 2R], [2R, -\infty]\}$

Table 1: Minimax Information States

$k = 0$	$W_{3,3}([0, 3C]) = 3C, W_{3,3}([2C + R, -\infty]) = 2C + R,$ $W_{3,3}([R + C, R + C]) = R + C, W_{3,3}([2R + C, -\infty]) = 2R + C,$ $W_{3,3}([R, 2R]) = 2R, W_{3,3}([2R, -\infty]) = 2R,$ $W_{3,3}([2R, 2R]) = 2R, W_{3,3}([3R, -\infty]) = 3R$
$k = 1$	$W_{2,3}([0, 2C]) = 3C$ and $\mu_2^\star([0, 2C]) = 0$ $W_{2,3}([R + C, -\infty]) = R + C$ and $\mu_2^\star([R + C, -\infty]) = 0$ $W_{2,3}([R, R]) = R + C$ and $\mu_2^\star([R, R]) = 0$ $W_{2,3}([2R, -\infty]) = 2R$ and $\mu_2^\star([2R, -\infty)] = 0$
$k = 2$	$W_{1,3}([0, C]) = \min\{3C, R + C\}$ and $\mu_1^\star([0, C]) = 0 \Leftrightarrow 2C < R$ $W_{1,3}([R, -\infty]) = R + C$ and $\mu_1^\star([R, -\infty]) = 0$
$k = 3$	$W_{0,3}([0, 0]) = \min\{3C, R + C\}$ and $\mu_0^\star([0, C]) = 0 \Leftrightarrow 2C < R$

Table 2: Dynamic Programming

2. $\pi_0 = [1, 0] \Rightarrow s_0 = [0, -\infty]$.

3. $\pi_0 = [0, 1] \Rightarrow s_0 = [-\infty, 0]$.

Let us consider the first class, corresponding to $s_0 = [0, 0]$. The first step is to generate all information states of interest using (15), beginning with $s_0 = [0, 0]$. The result is shown in Table 1. Next, we use the dynamic programming equations (20), (21) to determine the value function and the optimal control for each information state of interest. Let us denote the optimal policy by $\mu^\star$. The result is shown in Table 2.

We can proceed similarly for the other two classes of initial distributions, corresponding to $s_0 = [0, -\infty]$ and $s_0 = [-\infty, 0]$. The optimal policy can be described succinctly as follows. At $k = 2$ (one step from the end), do nothing. For $k < 2$, do as follows:

- If there is no possibility that the system is in the broken state, do nothing.
- Otherwise, do nothing if and only if $2C < R$.

Note that the policy does not depend on the values of θ and q, other than to the extent that $\theta > 0$ and $0 < q < 1$. This is consistent with our earlier remark that probabilities of system trajectories are significant only to the extent that they are zero or nonzero. The policy can indeed be interpreted as minimizing the worst case cost incurred in the system's evolution. If $2C < R$, then the optimal minimax policy will be never to repair the system, and thereby incur in the worst case a cost of C at each time. Alternatively, in the worst case, if we repair the system when it is possibly in the broken state, it will return to the broken state after one unit of time in the working state. Thus, again in the worst case we incur an average cost of $\frac{R}{2}$, which is greater than C.

5.2 Infinite Horizon

Let us assume that the state of the system is fully observed, i.e. $q = 1$. We wish to compare risk-neutral policies with risk-sensitive and minimax policies, and develop some intuition on what is the effect of increasing the risk-sensitivity parameter γ. In the average cost setting, this type of question has been addressed recently in [HHMF97]. In the discounted cost setting, comparisons are difficult due to non-stationarity of the optimal policies. Thus, we will use the generalized risk-sensitive and minimax formulations, with $\beta = \beta'' = 1$ (Porteus formulation).

For each of the criteria of interest, it turns out that the optimal policy is one of the following.

- policy μ_0, given by $\mu_0(0) = 0, \mu_0(1) = 0$. This is the "no action" policy, leading to the following risk-neutral and minimax value functions:

$$h^0_{\mu_0} = \begin{bmatrix} \frac{\beta'\theta C}{[1-\beta'(1-\theta)](1-\beta')} \\ \frac{C}{1-\beta'} \end{bmatrix}, \bar{h}_{\mu_0} = \begin{bmatrix} \frac{\beta' C}{(1-\beta')} \\ \frac{C}{1-\beta'} \end{bmatrix}. \tag{50}$$

- policy μ_1, given by $\mu_1(0) = 0, \mu_1(1) = 1$. This is the "repair when broken" policy, leading to the following risk-neutral and minimax value functions:

$$h^0_{\mu_1} = \begin{bmatrix} \frac{\beta'\theta R}{1-\beta(1-\theta+\beta'\theta)} \\ \frac{[1-\beta'(1-\theta)]R}{1-\beta'(1-\theta+\beta'\theta)} \end{bmatrix}, \bar{h}_{\mu_1} = \begin{bmatrix} \frac{\beta' R}{1-(\beta')^2} \\ \frac{R}{1-(\beta')^2} \end{bmatrix}. \tag{51}$$

The risk-sensitive value functions involve a policy evaluation iteration and cannot be represented analytically.

Comparing the value functions under the two policies of interest, we find that optimal risk-neutral and decision-making are characterized by a threshold value for $\frac{R}{C}$. Specifically, we have:

$$\begin{aligned} R < C \cdot t_{rn} \quad &\Leftrightarrow \quad \mu_1 \text{ is the optimal risk-neutral policy,} \\ R < C \cdot t_{mm} \quad &\Leftrightarrow \quad \mu_1 \text{ is the optimal minimax policy,} \end{aligned}$$

where

$$t_{rn} = \frac{1-\beta'(1-\theta+\beta'\theta)}{[1-\beta'(1-\theta)](1-\beta')}, \; t_{mm} = 1+\beta'. \tag{52}$$

t_{rn} and t_{mm} are the risk-neutral and the minimax thresholds, respectively. Since $t_{rn} - t_{mm} > 0$ for $0 < \beta', \theta < 1$, we conclude that for this system the risk-neutral

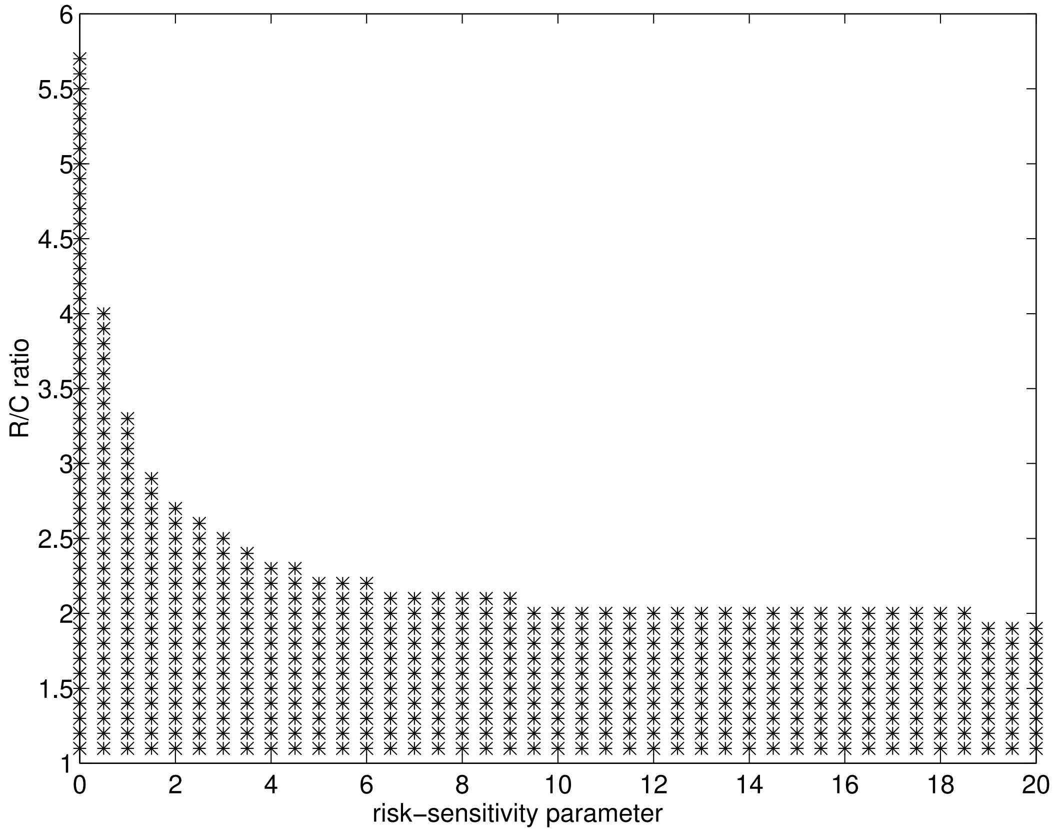

Fig. 1: Risk-sensitive threshold as a function of γ, for $\beta' = 0.9, \theta = 0.1$.

controller is more *aggressive* than the minimax controller, in that there is a larger range of values of $\frac{R}{C}$ for which a faulty unit is replaced.

The threshold value for the risk-sensitive criterion depends on the value for the risk-sensitivity parameter γ, and must be determined numerically. The results of our computations seem to confirm that, for all values of $0 < \beta', \theta < 1$, the value of the risk-sensitive threshold decreases as the risk-sensitivity parameter is increased. Figure 1 indicates results for a particular choice of β' and θ. Each asterisk in the plot indicates that, for the corresponding values of $\frac{R}{C}$ and γ, the optimal risk-sensitive policy is μ_1. Thus, the plot illustrates numerically determined risk-sensitive threshold values as a function of the risk-sensitivity parameter.

6 Conclusions

This paper has provided a number of contributions to the literature on risk-sensitive and minimax control for finite state systems. Key results include a large-risk-limit connection between risk-sensitive and minimax control in the MDP setting, infinite horizon discounted dynamic programming equations for both risk-sensitive and minimax criteria, and a generalized framework for discounted optimal decision-making, allowing for controllers that retain risk-sensitivity without sacrificing stationarity on the infinite horizon.

Acknowledgement This research was partially supported by the National Science Foundation under Grant EEC 9402384.

References

[BB95] T. Basar and P. Bernhard. *H^∞-Optimal Control and Related Minimax Design Problems.* Birkhauser, 1995.

[BJam] J. S. Baras and M. R. James. Robust and risk-sensitive output feedback control for finite state machines and hidden markov models. *Journal of Mathematical Systems, Estimation, and Control*, to appear.

[BR71] D. P. Bertsekas and I. B. Rhodes. On the minimax feedback control of uncertain systems. In *Proc. IEEE Conference on Decision and Control*, pages 451–455, 1971.

[BS85] A. Bensoussan and J. H. Van Schuppen. Optimal control of partially observable stochastic systems with an exponential-of-integral performance index. *SIAM Journal on Control and Optimization*, 23(4):599–613, 1985.

[Cor97] S. P. Coraluppi. *Optimal Control of Markov Decision Processes for Performance and Robustness.* PhD thesis, University of Maryland, 1997.

[CS87] K. J. Chung and M. J. Sobel. Discounted mdp's: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25:49–62, 1987.

[Eag75] J. N. Eagle. *A Utility Criterion for the Markov Decision Process.* PhD thesis, Stanford University, 1975.

[EZ89] L. G. Epstein and S. E. Zin. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica*, 57(4):937–969, 1989.

[FGMar] E. Fernández-Gaucherand and S. I. Marcus. Risk-sensitive optimal control of hidden markov models: Structural results. *IEEE Transactions on Automatic Control*, to appear.

[FHH(1)] W. H. Fleming and D. Hernández-Hernández. Risk-sensitive control of finite state machines on an infinite horizon I. *SIAM Journal on Control and Optimization*, to appear.

[FHH(2)] 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.

[GD88] 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 Letters*, 11:167–172, 1988.

[HHM96] D. Hernández-Hernández and S. I. Marcus. Risk-sensitive control of markov processes in countable state space. *Systems and Control Letters*, 29:147–155, 1996.

[HHM97] D. Hernández-Hernández and S. I. Marcus. Existence of risk sensitive optimal stationary policies for controlled markov processes. Technical report, University of Maryland, 1997.

[HHMF97] D. Hernández-Hernández, S. I. Marcus, and P. J. Fard. Analysis of a risk sensitive control problem for hidden markov chains. Technical report, University of Maryland at College Park, 1997.

[HM72] R. A. Howard and J. E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972.

[HS84] D. P. Heyman and M. J. Sobel. *Stochastic Models in Operations Research, Vol. II: Stochastic Optimization*. McGraw-Hill, 1984.

[HS95] L. P. Hansen and T. J. Sargent. Discounted linear exponential quadratic gaussian control. *IEEE Transactions on Automatic Control*, 40:968–971, 1995.

[Jac73] D. H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control*, 18(2):124–131, 1973.

[KP78] D. M. Kreps and E. L. Porteus. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*, 46(1):185–200, 1978.

[Lov89] W. S. Lovejoy. A note on exact solution of partially observed markov decision processes. Technical report, Graduate School of Business, Stanford University, 1989.

[McE96a] W. M. McEneaney. Risk-sensitive control of nonlinear systems. *SIAM Activity Group on Control and System Theory Newsletter*, 4(1), 1996.

[McE96b] W. M. McEneaney. Risk-sensitive control of nonlinear systems. *SIAM Activity Group on Control and System Theory Newsletter*, 4(2), 1996.

[Por75] E. Porteus. On the optimality of structured policies in countable stage decision processes. *Management Science*, 22(2):148–157, 1975.

[SS73] R. D. Smallwood and E. J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

[Whi81] P. Whittle. Risk-sensitive linear/quadratic/gaussian control. *Advances in Applied Probability*, 13:764–777, 1981.