Risk-Sensitive, Minimax, and Mixed Risk-Neutral/Minimax Control of Markov Decision Processes

by S.P. Coraluppi, S.I. Marcus

T.R. 98-30

Risk-Sensitive, Minimax, and Mixed Risk-Neutral/Minimax Control of Markov Decision Processes

Stefano P. Coraluppi
Steven I. Marcus

ABSTRACT: This paper analyzes a connection between risk-sensitive and minimax criteria for discrete-time, finite-state Markov Decision Processes (MDPs). We synthesize optimal policies with respect to both criteria, both for finite horizon and discounted infinite horizon problems. A generalized decision-making framework is introduced, leading to stationary risk-sensitive and minimax optimal policies on the infinite horizon with discounted costs.

We introduce the mixed risk-neutral/minimax objective, and utilize results from risk-neutral and minimax control to derive an information state process and dynamic programming equations for the value function. We synthesize optimal control laws both on the finite and infinite horizon, and establish the effectiveness of the controller as a tool to trade off risk-neutral and minimax objectives.

KEYWORDS: Markov Decision Processes, Risk-Sensitive Control, Minimax Control, Mixed Risk-Neutral/Minimax Control

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 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. Further connections between risk-sensitive control and a particular minimax control problem are explored in an interesting recent work (see [PJD97]).

An interesting fact both in risk-sensitive and minimax control is that in general, on the infinite horizon and with stationary discounted 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 [MFHCF97], [FHH(1)], [FHH(2)], [HHM96], [HHM97].

While risk-neutral and minimax controllers are limiting special cases of the risk-sensitive controller, we show that in general the risk-sensitive controller does not effectively trade off risk-neutral and minimax objectives. We introduce a mixed risk-neutral/minimax objective, solve the associated optimal control problem, and show that it does effectively trade off risk-neutral and minimax objectives, at the cost of increased controller complexity with respect to the risk-sensitive controller. Our mixed risk-neutral/minimax formulation parallels the mixed H_2/H_∞ criterion that has been introduced in the linear systems setting. See [ZGBD94] and [DZGB94] for details.

This paper is organized as follows. In Section 2 we discuss our results on risk-sensitive and minimax control. In Section 3, we define and address the mixed risk-neutral/minimax problem, and motivate its usefulness as a tool to trade off risk-neutral and minimax objectives. We note that throughout our presentation, proofs are omitted due to space constraints. For detailed proofs of our results and further discussion, the reader is referred to [Cor97].

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

$$J(\mu, \pi_0) = E^{\mu,\pi_0}[\sum_k c_k(x, u)], \tag{1.1}$$

where μ is a non-anticipative policy and π_0 is the probability distribution on the states of the system at time $k = 0$. A policy or control law is a sequence of mappings from available information to control actions. Let us denote by M this set of (non-anticipative) policies. Good references for the risk-neutral control of MDPs include [KV86], [Ber95], and [Put94].
If the state is observed, that is $y_k = x_k$, there exists a Markov policy that is optimal. In the partially observed setting, the conditional distribution of the state given past observations is an information state. It is defined recursively as follows:

$$\pi_{k+1} = r(\pi_k, u_k, y_{k+1}) = \frac{\pi_k P(u_k)\bar{Q}(y_{k+1}, u_k)}{\pi_k P(u_k)\bar{Q}(y_{k+1}, u_k)\underline{1}} \tag{1.2}$$

where $\bar{Q}(\cdot,\cdot)$ is a diagonal matrix with $\bar{Q}_{ii}(y, u) = pr(y_{k+1} = y|x_{k+1} = i, u_k = u)$, and $\underline{1} = [1, \ldots, 1]'$. The information state at each time k is a $|X|$-dimensional vector belonging to the space Π, the unit simplex in $\Re_+^{|X|}$, where $\Re_+$ is the set of non-negative real numbers. The value function has the important properties that it is piecewise linear and concave in π_k.
The *risk-sensitive* objective is given by

$$J^\gamma(\mu, \pi_0) = \frac{1}{\gamma} \log E^{\mu,\pi_0}[\exp(\gamma \sum_k c_k(x, u))]. \tag{1.3}$$

For small γ, (1.3) takes the form

$$J^\gamma(\mu, \pi_0) \simeq E^{\mu,\pi_0}[\sum_k c_k(x, u)] + \frac{\gamma}{2} Var^{\mu,\pi_0}[\sum_k c_k(x, u)], \tag{1.4}$$

and in the limit $\gamma \to 0$, (1.3) reverts to the risk-neutral objective (1.1). 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 (1.3) is given by

$$\hat{J}^{\gamma}(\mu, \pi_0) = E^{\mu,\pi_0}[\exp(\gamma \sum_k c_k(x,u))]. \tag{1.5}$$

In [BJam], an information state process for the MDP with respect to criterion (1.5) is defined, satisfying the following recursion:

$$\begin{aligned} \sigma_0^{\gamma} &= \pi_0, & (1.6)\\ \sigma_{k+1}^{\gamma} &= |Y|\sigma_k^{\gamma} D^{\gamma}(k, u_k)\bar{Q}(y_{k+1}, u_k), & (1.7) \end{aligned}$$

where

$$D_{ij}^{\gamma}(k,u) := P_{ij}(u)\exp(\gamma c_k(i,u)), \tag{1.8}$$

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

$$S_{k,N}^{\gamma}(\sigma) := \inf_{\mu \in M} E^{\dagger}[\sigma_N^{\gamma} \cdot \exp(\gamma c_N)|\sigma_k^{\gamma} = \sigma]. \tag{1.9}$$

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 (1.9) are given by

$$\begin{aligned} S_{N,N}^{\gamma}(\sigma^{\gamma}) &= \sigma^{\gamma} \cdot \exp(\gamma c_N), & (1.10)\\ S_{k,N}^{\gamma}(\sigma^{\gamma}) &= \min_{u \in U} E^{\dagger}[S_{k+1,N}^{\gamma}(|Y|\sigma^{\gamma}D^{\gamma}(k,u)\bar{Q}(y_{k+1},u)]. & (1.11) \end{aligned}$$

It has been shown in [FGMar] that $S_{k,N}^{\gamma}(\cdot)$ 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

$$\bar{J}(\mu, \pi_0) = \sup_{\omega \in \Omega^{\mu}} \sum_k c_k(x_k, u_k), \tag{1.12}$$

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.

2.1 Finite Horizon Results

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 for all $\omega \in \Omega$ we have $F(\omega) = \lim_{\gamma\to\infty} F^\gamma(\omega)$. Also, let $p(\omega)$ be a nonnegative real number $\forall \omega \in \Omega$, independent of γ. Then

$$\lim_{\gamma\to\infty} \frac{1}{\gamma} \log \sum_{\omega\in\Omega} p(\omega) \exp\left[\gamma F^\gamma(\omega)\right] = \max_{\omega\in\Omega, p(\omega)\neq 0} F(\omega). \quad \square \tag{1.13}$$

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

$$s_k := \lim_{\gamma\to\infty} \frac{1}{\gamma} \log \sigma_k^\gamma, \ \forall k, \tag{1.14}$$

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:

$$\begin{aligned} s_0[x] &= \begin{cases} 0 & \text{if } \pi_0[x] \neq 0, \\ -\infty & \text{otherwise}, \end{cases} && (1.15) \\ s_{k+1}[x'] &= f_k(s_k, u_k, y_{k+1}). && (1.16) \end{aligned}$$

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

$$f_k(s_k, u_k, y_{k+1}) = \begin{cases} \max_{x\in\tilde{X}(x',u_k)}[s_k[x] + c_k(x,u_k)] & \text{if } \begin{array}{l} \tilde{X}(x',u_k) \neq \emptyset, \\ x' \in \tilde{Y}(y_{k+1},u_k) \end{array} \\ -\infty & \text{otherwise}, \end{cases} \tag{1.17}$$

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 (1.12) on the finite horizon are related by the following:

$$\bar{J}(\mu, \pi_0) = \max_{y_1,\ldots,y_N} \max_{i\in X} s_N[i]. \tag{1.18}$$

This motivates the following definition for the value function:

$$W_{k,N}(s) := \min_{\mu\in M} \max_{y_{k+1},\ldots,y_N} \max_{i\in X} s_N[i], \text{ where } s_k = s. \tag{1.19}$$

Indeed, we have

$$W_{0,N}(s_0) = W_{0,N}(\lim_{\gamma\to\infty} \frac{1}{\gamma} \log \pi_0) = \min_{\mu\in M} \bar{J}(\mu, \pi_0). \qquad (1.20)$$

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 (1.15), (1.16) 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 (1.19). 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, $\forall s \in \tilde{R}_+^{|X|}$:

$$\begin{aligned} W_{N,N}(s) &= \max_{i\in X} s[i], & (1.21)\\ W_{k,N}(s) &= \min_{u\in U}\max_{y\in Y} W_{k+1,N}(f(s,u,y)). & (1.22)\end{aligned}$$

A policy that achieves the minimum in equations (1.21) and (1.22), also achieves the minimum in (1.19). Furthermore, the policy is separated and is optimal with respect to (1.12). □

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. See [Cor97] for details. Without such results, the minimization in (1.11) 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 (1.21), (1.22) to find the optimal control at each state of interest.

2.2 The Infinite Horizon Case

One way to insure that the objectives (1.1), (1.3), and (1.12) are bounded on the infinite horizon is to introduce 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

$$\hat{S}_k^\gamma(x) := \lim_{N\to\infty} \hat{S}_{k,N}^\gamma(x) \tag{1.23}$$

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

$$\hat{S}_0^\gamma = \min_{u\in U} \{D^\gamma(0,u)\hat{S}_0^{\beta\gamma}\}, \tag{1.24}$$

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

$$S_k^\gamma(\sigma) := \lim_{N\to\infty} S_{k,N}^\gamma(\sigma), \tag{1.25}$$

where $S_{k,N}^\gamma$ is defined in (1.9). The limit in (1.25) exists, and

$$S_0^\gamma(\sigma) = \min_{u\in U} E^\dagger[S_0^{\beta\gamma}(|Y|\sigma D^\gamma(0,u)\bar{Q}(y_1,u))]. \ \Box \tag{1.26}$$

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

$$W_k(s) := \lim_{N\to\infty} W_{k,N}(s). \tag{1.27}$$

We can verify that the limit in (1.27) is well-defined by recalling that $W_{k,N} = \lim_{\gamma\to\infty} \frac{1}{\gamma} \log S_{k,N}^\gamma(\exp(\gamma s))$, and $\lim_{N\to\infty} S_{k,N}$ is well-defined. Thus

$$W_k(s) = \lim_{\gamma\to\infty} \frac{1}{\gamma} \log S_k^\gamma(\exp(\gamma s)) \tag{1.28}$$

We can relate the value function to the criterion (1.12) by taking the limit in (1.20) as $N \to \infty$. We obtain:

$$W_0(s_0) = \inf_{\mu\in M} \bar{J}(\mu,\pi_0) \tag{1.29}$$

The following result characterizes the infinite horizon value function.

Theorem 3 (Minimax Infinite Horizon Dynamic Programming). The value function (1.27) satisfies the following, $\forall s \in \tilde{R}_+^{|X|}$:

$$W_0(s) = \min_{u\in U} \max_{y\in Y} \beta W_0(\frac{f_0(s,u,y)}{\beta}). \ \Box \tag{1.30}$$

In risk-neutral control, with finite state and action spaces, there exists a stationary optimal policy. In the full state observations setting, this 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 (1.26) and (1.30) 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 $\xi = \log[(1-\beta)\epsilon / \parallel c \parallel] / \log \beta$, and $\parallel c \parallel := \max_{x \in X, u \in U} |c(x,u)|$. 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 (1.3) and (1.12) are within ϵ of optimal. See ([Cor97]) for details.

2.3 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 (1.10), (1.11) or (1.21), (1.22) 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 (1.3) can be defined as

$$s^{\gamma}_{k,N}(i) := \min_{\mu} \frac{1}{\gamma} \log E^{\mu}[\exp(\gamma C_{k,N}) | x_k = i], \; i \in X \qquad (1.31)$$

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 (1.31) are given by

$$\begin{aligned} s^{\gamma}_{k,N}(i) &= \min_{u \in U}\{c_k(i,u) + \frac{1}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma s^{\gamma}_{k+1,N}(j))]\}, \quad (1.32) \\ s^{\gamma}_{N,N}(i) &= c_N(i). \quad (1.33) \end{aligned}$$

In the small-risk limit, $\gamma \to 0$, (1.32), (1.33) revert to the usual risk-neutral dynamic programming equations. On the infinite horizon, we have

$$s^{\gamma}_{k}(i) = \min_{u \in U}\{c_k(i,u) + \frac{1}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma s^{\gamma}_{k+1}(j))]\}, \; k = 0, \ldots . \quad (1.34)$$

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

$$s^{\gamma}_{k+1}(\cdot) = \beta s^{\beta\gamma}_{k}(\cdot). \tag{1.35}$$

Equation (1.35) also reverts to a well-known relationship in the risk-neutral case:

$$s^{0}_{k+1}(\cdot) = \beta s^{0}_{k}(\cdot). \tag{1.36}$$

A more general set of optimality equations than (1.32), (1.33) can be defined as follows:

$$\begin{aligned} h^{\gamma}_{k,N}(i) &= \min_{u\in U}\{c_k(i,u) + \\ &\qquad \frac{\beta'}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma\beta'' h^{\gamma}_{k+1,N}(j))]\}, \qquad (1.37) \\ h^{\gamma}_{N,N}(i) &= c_N(i). \qquad (1.38) \end{aligned}$$

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

$$h^{\gamma}_{k}(i) = \min_{u\in U}\{\beta^k c(i,u) + \frac{\beta'}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma\beta'' h^{\gamma}_{k+1}(j))]\},\ k = 0,\ldots. \tag{1.39}$$

Once again we obtain the classical, Porteus, and Eagle formulations as special cases of (1.39). A key feature of the generalized formulation (1.39) 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^{γ}_{k}. Thus, by setting either β' or β'' to be less than one, we can set $\beta = 1$. It can then be shown that $h^{\gamma}_{k}(\cdot) = h^{\gamma}(\cdot)$, that is we have a time-invariant value function, and furthermore there is a stationary policy that achieves the minimum in (1.16). 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 (1.39). We obtain the following:

$$h^{0}_{k}(i) = \min_{u\in U}\{\beta^k c(i,u) + \beta'\beta'' \sum_j P_{ij}(u) h^{0}_{k+1}(j)\},\ k = 0,\ldots. \tag{1.40}$$

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 (1.39) 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

$$\begin{aligned} \bar{h}_{k,N}(i) &= \min_{u \in U}\{c_k(i,u) + \beta'\beta'' \max_{j \in \tilde{X}'(i,u)} \bar{h}_{k+1,N}(j)\}, && (1.41) \\ \bar{h}_{N,N}(i) &= c_N(i), && (1.42) \end{aligned}$$

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

$$\bar{h}_k(i) = \min_{u \in U}\{\beta^k c(i,u) + \beta'\beta'' \max_{j \in \tilde{X}'(i,u)} \bar{h}_{k+1}(j)\}. \qquad (1.43)$$

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 (1.32), (1.33), the equations (1.37), (1.38) 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 (1.37), (1.38) 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 (1.39). This can be understood in light of the axiomatic foundation of Utility Theory (see e.g. [HS84]), and some dynamic extensions discussed in [KP78].

3 Mixed Risk-Neutral/Minimax Control

The approach for defining the mixed risk-neutral/minimax objective is the following. We let a bound be given on the worst-case cost incurred, as a function of the probability distribution on $x_0 \in X$. Subject to this bound, an optimal policy is one for which the expected cost incurred is minimized. Specifically, let $\eta(\cdot) : \tilde{\Re}_+^{|X|} \to \Re_+$ be given, such that $\eta(s_0) \geq \eta_0(s_0)$, $\forall s_0 \in \tilde{\Re}_+^{|X|}$, where $\eta_0(s_0) = \min_{\mu \in M} \bar{J}(\mu, s_0)$. Recall that s_0 depends on π_0 as given by (1.15). Given the functional dependence of $\bar{J}(\mu,\pi_0)$ on π_0, by a slight abuse of notation we may write $\bar{J}(\mu, s_0)$ instead.

We define $M(\eta(\cdot)) \subset M$ to be the set of feasible policies such that for each initial probability distribution $\pi_0 \in \Pi$, where Π denotes the unit simplex, the worst-case cost incurred does not exceed $\eta(s_0)$. That is, for $\mu \in M(\eta(\cdot))$ and for $s_0 \in \tilde{\Re}_+^{|X|}$,

$$\bar{J}(\mu, s_0) = \max_{\omega \in \Omega_0, p^{\mu,\pi_0}(\omega) \neq 0} \sum_k c_k(x, u)(\omega) \leq \eta(s_0). \tag{1.44}$$

We seek a policy $\mu^\star$ that minimizes the risk-neutral objective subject to a constraint on the allowable worst-case cost. That is, given $\eta(\cdot) \geq \eta_0(\cdot)$, an optimal policy $\mu^\star$ is one which satisfies

$$J(\mu^\star, \pi_0) = \min_{\mu \in M(\eta(\cdot))} J(\mu, \pi_0), \tag{1.45}$$

for all $\pi_0 \in \Pi$. Again, recall that s_0 depends on π_0 as given by (1.15).

3.1 Finite Horizon Results

In the general, partially observed setting, we wish now to address the task of determining an optimal policy $\mu^\star$ as defined by (1.45), for a given $\eta(\cdot) > \eta_0(\cdot)$. We will need to introduce an appropriate sufficient statistic, as well as dynamic programming equations for the value function.

We introduce the following statistic which combines the risk-neutral and the minimax information states. This statistic will be our candidate information state (sufficient statistic). This statistic is given by $\{g_k\}, k = 0, 1, \ldots$, where $g_k := (\pi_k, s_k)$.

We now introduce a number of definitions. Let Ω_k be the set of trajectories of the system beginning at time k. That is, elements of Ω_k are of the form $(x_k, u_k, x_{k+1}, \ldots)$. Let $p^{\mu,g,k}(\omega), \omega \in \Omega_k$, denote the probability of trajectory ω given that the information state at time k is $g_k = g$. Let $M(\eta(s_0), g, k) \subset M$ be the set of policies such that

$$\max_{\omega \in \Omega_k, p^{\mu,g,k}(\omega) \neq 0} [s[x_k] + \sum_{l=k}^{N} c_l(x_l, u_l)](\omega) \leq \eta(s_0). \tag{1.46}$$

That is, a policy μ is in $M(\eta(s_0), g, k)$ if the worst case cost incurred, given that the information state is g at time k, is no greater than $\eta(s_0)$. We say that an information state g is *feasible* at time k with respect to $\eta(s_0)$ if $G(\eta(s_0), g, k) \neq \emptyset$. Let $U(\eta(s_0), g, k) \subset U$ be the set of feasible controls, that is for $u \in U(\eta(s_0), g, k)$, $\exists \mu \in M(\eta(s_0), g, k)$ such that $\mu_k(g) = u$.

Define the value function $V^{\eta(s_0)}$ as follows:

$$V_{k,N}^{\eta(s_0)}(g) := \min_{\mu \in M(\eta, g, k)} E[\sum_{l=k}^{N} c_l(x_l, u_l) | g_k = g]. \tag{1.47}$$

In particular, we have

$$V_{0,N}^{\eta(s_0)}(g_0) = \min_{\mu \in M(\eta(s_0),g_0,0)} J(\mu,\pi_0) = J(\mu^\star,\pi_0), \qquad (1.48)$$

using equation (1.45).

Theorem 4 (Dynamic Programming). The value function defined in equation (1.47) satisfies the following dynamic programming equations for all feasible g:

$$\begin{aligned} V_{N,N}^{\eta(s_0)}(g) &= \pi \cdot c_N, & (1.49)\\ V_{k,N}^{\eta(s_0)}(g) &= \min_{u \in U(\eta(s_0),g,k)} E[c_k(x,u) + V_{k+1,N}^{\eta(s_0)}(g_{k+1}) | g_k = g]. & (1.50)\end{aligned}$$

Furthermore, a policy $\mu_s^\star$ that achieves the minimum in equations (1.49) and (1.50) also achieves the minimum in (1.47). The optimal separated policy $\mu_s^\star$ is optimal within the larger class $M(\eta(s_0),(\pi_0,s_0),0)$ of all feasible policies. □

Note that for a given time k, the feasible information states g for which we are interested in the minimization in (1.50) will be uncountably infinite in general. Thus we need structural results for the value function to make the minimization tractable. The following two lemmas will be useful to address this.

Lemma 2 [Ast69]. Let $f_1(x)$ and $f_2(x)$ be concave functions. The function $f(x) = \min\{f_1(x), f_2(x)\}$ is also concave. □

Lemma 3 [Ast69]. Let the function $g : \Pi \to \Re$ be concave and let A be a linear transformation from Π into Π. Then the function $f : \Pi \to \Re$ defined by

$$f(x) = \| Ax \| \cdot g(\frac{Ax}{\| Ax \|}), \ x \in \Pi, \qquad (1.51)$$

is also concave. □

Using these two lemmas, we can show the following.

Theorem 5 (Concavity). The value function $V_{k,N}^{\eta(s_0)}(g) = V_{k,N}^{\eta(s_0)}(\pi,s)$ is concave as a function of π. □

Theorem 6 (Piecewise Linearity). The value function $V_{k,N}^{\eta(s_0)}(g) = V_{k,N}^{\eta(s_0)}(\pi,s)$ is piecewise linear as a function of π. □

The determination of optimal policies on the finite horizon can be achieved by generalizing the methodology used for risk-neutral control. A key observation is that only a finite number of values of the minimax information state will be of interest. Thus a scheme for determining optimal finite horizon policies is the following:

1. Generate all minimax information states s_k of interest, for $k = 0, 1, \ldots$ Discard those information states such that the corresponding g will be infeasible.

2. Implement a backwards dynamic programming iteration using (1.49), (1.50). For each k, $0 \leq k < N$, we must consider states $g = (\pi, s)$ such that s is generated by step (1) and $\pi \in \Pi$. For each value s, a risk-neutral methodology can be utilized.

In the worst case, the number of minimax information states will increase polynomially in the size of the horizon as follows:

$$|s_k| = (2^{|X|} - 1)(|U| \cdot |Y|)^k. \tag{1.52}$$

Also, in the worst case, the number of vectors needed to represent the value function $V_{k,N}(s, \cdot)$ is given by

$$|U|^{(|Y|^{N-k}-1)/(|Y|-1)}. \tag{1.53}$$

This can be derived by noting that the number of vectors needed at time k, or $|A_k|$, increases as follows:

$$|A_k| \leq |U| \cdot |A_{k+1}|^{|Y|}. \tag{1.54}$$

Thus, the controller complexity at time k is bounded by the product of (1.52) and (1.53).

A slight reduction in the complexity of the algorithm can be obtained with the following observation. Our algorithm is such that at time k, we consider separately information states $g_k = (\pi, s)$ and $g'_k = (\pi, s')$, with corresponding bounds given by $\eta(s_0)$ and $\eta(s'_0)$, respectively. Note that if $\eta(s_0)\underline{1} - s_0 = \eta(s'_0)\underline{1} - s'_0$, we need not repeat the minimization in (1.50) both for $s_k = s$ and $s_k = s'$. This observation leads to a more efficient procedure to determine an optimal policy in many instances, though the worst-case complexity is the same.

In the special case where the state of the system is observed, we know that both in risk-neutral and in minimax control there is a Markov policy that is optimal. Unfortunately this is not the case for the mixed control problem. The information state process $g_k, k = 0, 1, \ldots$, cannot be simplified in this manner. Intuitively, this follows from the fact that at time k, the optimal policy depends not only on the state of the system but on the total accumulated cost up to time k.

The complexity of the mixed risk-neutral/minimax controller is greater than that of the risk-sensitive controller in general. In the fully observed setting, it is well known that Markov policies are optimal for the risk-sensitive criterion (see [HM72]). In the general, partially observed setting, it has been shown (see [Cor97]) that the complexity of the risk-sensitive controller is the same as the risk-neutral controller.

3.2 The Infinite Horizon Case

We derive dynamic programming equations characterizing the value function on the infinite horizon. Defining $V_k^{\eta(s_0)} := \lim_{N\to\infty} V_{k,N}^{\eta(s_0)}$, we obtain the following, using (1.50):

$$V_k^{\eta(s_0)}(g) = \min_{u\in U(\eta(s_0),g,k)} E[c_k(x,u) + V_{k+1}^{\eta(s_0)}(g_{k+1})]. \tag{1.55}$$

Assuming $c_k(\cdot,\cdot) = \beta^k c(\cdot,\cdot)$, time-shifted value functions can be related by observing that

$$V_{k+1}^{\eta(s_0)}(\pi, s) = \beta V_k^{\frac{\eta(s_0)}{\beta}}(\pi, \frac{s}{\beta}). \tag{1.56}$$

Combining (1.55) and (1.56), we obtain the following equation characterizing the value function:

$$\begin{aligned} V_k^{\eta(s_0)}(\pi, s) \quad = \quad & \min_{u\in U(\eta(s_0),(s,\pi),0)} E[c(x,u) + \\ & \beta V_k^{\frac{\eta(s_0)}{\beta}}(r(\pi, u, y_{k+1}), \frac{f_k(s,u,y_{k+1})}{\beta})]. \end{aligned} \tag{1.57}$$

This equation reverts to the risk-neutral dynamic programming equation as $\eta(s_0) \to \infty$, that is as we relax the constraint on worst-case cost.
On the infinite horizon the optimal policy will be non-stationary in general, as with the minimax control problem. This fact makes it difficult to directly utilize equation (1.57) in constructing an optimal policy. A near-optimal policy can be determined by considering an appropriate finite horizon approximation, as established by the following result.

Theorem 7 (Finite Horizon Approximation). Consider the MDP on the infinite horizon, with initial distribution π_0 on the states, and $\eta(\cdot) > \eta_0(\cdot)$. Let $\epsilon > 0$ be given. Then $\exists N > 0$ such that the policy $\hat{g}$ satisfies the following, $\forall \pi_0 \in \Pi$:

$$J(\hat{\mu}, \pi_0) - J(\mu^\star, \pi_0) < \epsilon, \tag{1.58}$$

$$\bar{J}(\hat{\mu}, \pi_0) < \eta(s_0) + \epsilon, \tag{1.59}$$

where $\mu^\star$ is an optimal mixed risk-neutral/minimax policy with robustness bound $\eta(\cdot)$, and $\hat{\mu}$ is an optimal mixed risk-neutral/minimax finite horizon policy on $(0, N-1)$, with $c_N = 0$ and robustness bound $\eta(\cdot)$. The policy $\hat{\mu}$ on $(N, \ldots)$ is arbitrary but fixed. □

In general, it is not possible to construct a near-optimal policy through a finite horizon approximation if we require the worst-case cost to be no greater than $\eta(s_0)$. That is, it is necessary to relax the bound on worst-case cost by ϵ, in order to achieve near-optimality in performance.

3.3 Control for Performance and Robustness

We will refer to the risk-neutral objective, which indicates expected total cost incurred, as a system's *performance*. Also, we will refer to the minimax objective, which indicates worst-case total cost incurred, as a system's *robustness*. When both objectives are of interest, we would like to utilize a family of controllers that provides a good way to trade off performance and robustness. In this section we will quantify what we mean by "good", and we will examine both the mixed risk-neutral/minimax and the risk-sensitive families of controllers in this light.

It is easy to see that the mixed risk-neutral/minimax controller has risk-neutral and minimax controllers as limiting cases. Specifically, as $\eta(s_0) \rightarrow \infty$ we have $M(\eta(s_0), g_0, 0) \rightarrow M$, so that $\lim_{\eta(s_0)\rightarrow\infty} V_{0,N}^{\eta(s_0)}(g_0) = V_{0,N}(\pi_0)$ using (1.48). That is, as we relax the constraint on worst-case behavior we recover the risk-neutral formulation. Similarly, as $\eta(s_0) \rightarrow \eta_0(s_0)$, the mixed risk-neutral/minimax controller will be an optimal minimax controller. In general, there may be more than one minimax controller, since there may be more than one policy achieving the robustness bound $\eta_0(s_0)$.

As noted earlier, the risk-sensitive controller also has risk-neutral and minimax controllers as limiting cases, as $\gamma \rightarrow 0$ and $\gamma \rightarrow \infty$, respectively.

While both families of controllers provide a link between the risk-neutral and minimax objectives, this itself is not sufficient to motivate the use of either family to trade off performance and robustness. Additional properties of the families of controllers are required. We proceed by first introducing some terminology. For the purposes of this discussion, we will not distinguish between two policies for which the performance (1.1) and the robustness (1.12) are the same. The terminology that we introduce in this section is in part borrowed from the language of portfolio theory. See [Sha70] for details.

We say that a policy μ *dominates* another policy μ' if the performance and robustness characteristics of μ are both at least as good as those of μ', for all probability distributions $\pi_0 \in \Pi$ on the initial state x_0. We say that a policy is *efficient* if it is dominated by no policy other than itself. We say that a family of policies is *efficient* if each policy in the family is itself an efficient policy. We say that a family of policies is *complete* if it is efficient, and if every efficient policy belongs to the family. We say that a family of policies is *monotonic* in a parameter if, for each probability distributions

$\pi_0 \in \Pi$ on the initial state x_0, a decrease (increase) in the parameter does not worsen performance, and an increase (decrease) in the parameter does not worsen robustness.

In order to effectively determine a policy which trades off performance and robustness as desired, one would like to identify a family of policies indexed by a parameter, that is both monotonic in the parameter and efficient. Then, one can search among this class of efficient policies, adjusting the parameter in a straighforward manner. If the family is also complete, one can achieve a more precise tradeoff than if it is not.

In the family of all risk-neutral optimal policies, there is exactly one that is efficient, the policy for which criterion (1.12) is smallest. Likewise, in the family of all minimax optimal policies, there is exactly one efficient policy, the policy for which criterion (1.1) is smallest. Other policies in these families, if they exist, are not efficient, though they are not dominated by any policy not in the respective family. Clearly then, the family of all risk-neutral optimal policies is only efficient if it consists of a single policy. The same is true of the family of all minimax optimal policies. Both families are complete if and only if there is a unique risk-neutral optimal policy, a unique minimax optimal policy, and these are the same.

By construction, for a given $\eta(\cdot) \geq \eta_0(\cdot)$, the mixed risk-neutral/minimax optimal policy is efficient. It follows immediately that the family of all mixed risk-neutral/minimax policies, $\{\mu^\eta, \eta(\cdot) \geq \eta_0(\cdot)\}$, is efficient. Furthermore, the family is complete. Indeed, let μ be any efficient policy, and let $\eta(\cdot)$ be its corresponding robustness. Since there is a mixed risk-neutral/minimax policy with threshold $\eta(\cdot)$, it follows that μ must be a mixed risk-neutral/minimax optimal policy. Finally, the family is monotonic in $\eta(\cdot)$. Indeed, as we increase η, we degrade the robustness characteristics and monotonically improve performance. This follows by observing (1.50) and noting that for $\eta_2(s_0) > \eta_1(s_0)$, $U(\eta_1(s_0), g, k) \subset U(\eta_2(s_0), g, k)$, $\forall g, k$. It follows that $V_{k,N}^{\eta_2(s_0)}(g) \leq V_{k,N}^{\eta_1(s_0)}(g)$.

Efficient policies are deterministic. Note that since the family of mixed risk-neutral/minimax optimal policies is a (complete) family of deterministic policies, it follows that every efficient policy is deterministic. Another simple way that this property of an efficient policy can be established is the following. Let μ_{nd} be a non-deterministic policy whereby with probability p we choose the (deterministic) policy μ_{d1}, and with probability $(1-p)$ we choose the (deterministic) policy μ_{d2}, $\mu_{d2} \neq \mu_{d1}$. We will show that μ_{nd} is not efficient. Since every non-deterministic policy can be expressed as a convex combination of deterministic policies, we will conclude that every efficient policy is deterministic. Let $\pi_0 \in \Pi$ be given. Let the performance under the two deterministic policies be p_{d1} and p_{d2} respectively, and let the robustness be r_{d1} and r_{d2}. The worst-case cost incurred under policy μ_{nd} will equal the greater of that for μ_{d1} and for μ_{d2}. That is,

$$r_{nd} = \max\{r_{d1}, r_{d2}\}. \tag{1.60}$$

The performance under policy μ_{nd} is given by

$$p_{nd} = p \cdot p_{d1} + (1-p) \cdot p_{d2}. \tag{1.61}$$

If $p_{d1} < p_{d2}$, we have $p_{d1} < p_{nd}$ and $r_{d1} \leq r_{nd}$, so that μ_{nd} is dominated by μ_1 and so is not efficient. Similarly, if $p_{d2} < p_{d1}$, we have $p_{d2} < p_{nd}$ and $r_{d2} \leq r_{nd}$, so that μ_{nd} is dominated by μ_1 and so is not efficient. If $p_{d1} = p_{d2}$, then since $\mu_{d1} \neq \mu_{d2}$, it must be that $r_{d1} < r_{d2}$ or $r_{d1} < r_{d2}$. Assume *w.l.o.g.* that $r_{d1} < r_{d2}$. Then $p_{d1} = p_{nd}$ and $r_{d1} < r_{nd}$, so again we conclude that μ_{nd} is not efficient. We conclude that every efficient policy is deterministic.

A Risk-Sensitive Example. The following example shows that, in general, the family $\{\mu^\gamma, \gamma > 0\}$ of risk-sensitive controllers is not efficient, and is not monotonic in γ. Consider a fully observed MDP evolving on a horizon of size $N = 1$, with state space $X = \{1,2,3\}$, and control space $U = \{1,2,3\}$. Let the probability transition matrices $P(u), u \in U$ be given by

$$P(u) = \begin{bmatrix} 0.5-\theta(u) & 2\theta(u) & 0.5-\theta(u) \\ 0.5-\theta(u) & 2\theta(u) & 0.5-\theta(u) \\ 0.5-\theta(u) & 2\theta(u) & 0.5-\theta(u) \end{bmatrix}, \tag{1.62}$$

where $0 \leq \theta(u) \leq 0.5, u \in U$. Let the cost at time 0 be given by $c_0(x,1) = 0$, $c_0(x,2) = \xi$, $c_0(x,3) = 2\xi$, with $\xi > 0$, for $x \in X$. Let the terminal cost at time 1 be given by $c_1(1) = 0$, $c_1(2) = c$, and $c_1(3) = 2c$. In particular, set $\xi = 0.01$, $c = 1$, $\theta(1) = 0$, $\theta(2) = 0.49$, $\theta(3) = 0.5$.

It is easy to verify the following. The risk-neutral policy is to select action $u = 1$ at time 0, for any initial states $x \in X$. The minimax policy is to select action $u = 3$ at time 0, for any initial states $x \in X$. For $\gamma = 0.1$, the risk-sensitive policy is to select action $u = 2$ at time 0, for any initial states $x \in X$. The risk-sensitive policy with $\gamma = 0.1$ is dominated by the risk-neutral policy, showing that the family of risk-sensitive policies is not efficient and is not monotonic in γ.

4 Conclusions

This paper overviews 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.

In addition, the paper discusses a mixed risk-neutral/minimax objective. The optimal control problem is addressed by generalizing known results for risk-neutral and minimax control. On the infinite horizon, ϵ-optimal policies are constructed by considering a sufficiently large, finite horizon approximation. The mixed risk-neutral/minimax objective provides a family of controllers that can be used to effectively trade off performance and robustness in controller design.

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

References

[Ast69] K. Astrom. Optimal control of markov processes with incomplete state information ii. the convexity of the loss function. *Journal of Mathematical Analysis and Applications*, 26:403–406, 1969.

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

[Ber95] D. P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 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*, 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.

[DZGB94] J. Doyle, K. Zhou, K. Glover, and B. Bodenheimer. Mixed H_2 and H_∞ performance objectives II: optimal control. *IEEE Transactions on Automatic Control*, 39(8):1575–1587, 1994.

[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*, 42(10): 1418-1422, 1997.

[FHH(1)] 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.

[FHH(2)] 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*, 35(5): 1790-1810, 1997.

[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. *Applied Mathematics and Optimization*, to appear.

[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.

[KV86] P. R. Kumar and P. Varaiya. *Stochastic Systems: Estimation, Identification, and Adaptive Control.* Prentice-Hall, 1986.

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

[MFHCF97] S. Marcus, E. Fernandez-Gaucherand, D. Hernandez-Hernandez, S. Coraluppi, P. Fard. Risk-sensitive markov decision processes. *Systems and Control in the Twenty-First Century*, 263-279. C. I. Byrnes, et. al. (eds.), Birkhauser, 1997.

[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.

[PJD97] I. Peterson, M. James, and P. Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. In *Proc. IEEE Conference on Decision and Control*, San Diego, CA, Dec. 1997.

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

[Put94] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley and Sons, 1994.

[Sha70] W. Sharpe. *Portfolio Theory and Capital Markets.* McGraw-Hill, 1970.

[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.

[ZGBD94] K. Zhou, K. Glover, B. Bodenheimer, and J. Doyle. Mixed H_2 and H_∞ performance objectives I: robust performance analysis. *IEEE Transactions on Automatic Control*, 39(8):1564–1574, 1994.

ALPHATECH, Inc., 50 Mall Road, Burlington, Massachusetts 01803, U.S.A.;
E-mail stefano.coraluppi@alphatech.com

Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, Maryland 20742, U.S.A.; E-mail marcus@isr.umd.edu