Mixed Risk-Neutral/Minimax Control of Discrete-Time, Finite-State Markov Decision Processes*

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

Abstract

This paper addresses the control design problem for discrete-time, finite-state Markov Decision Processes (MDPs), when both risk-neutral and minimax objectives are of interest. 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. We study the effectiveness of both the mixed risk-neutral/minimax family and the risk-sensitive family of controllers as tools to trade off risk-neutral and minimax objectives. We conclude that the mixed risk-neutral/minimax family is more effective, at the cost of increased controller complexity.

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

1 Introduction

The traditional approach to control system design for stochastic systems has been to consider a criterion to be minimized that is the expected cost incurred in the system's evolution. This approach is referred to as risk-neutral control, since the variance and higher order moments in the probability distribution of the cost are neglected. An alternative approach is to seek to minimize the worst case behavior in the system's evolution. This approach is referred to as minimax control, and is pursued in [6] in the partially observed Markov Decision Process (MDP) setting. In this work we will be interested in studying a formulation of the optimal control problem for MDPs which balances the objectives of minimizing expected cost incurred and worst case cost incurred.

One generalization of the risk-neutral approach is known as risk-sensitive control, whereby we seek to minimize a criterion that depends on the expected cost incurred as well as on higher order moments. Early formulations are due to [11] in the context of linear systems, and [10] in the MDP setting. Key results in the partially observed setting can be found in [3] and [20] for linear systems, and in [2], [8], and [6] for MDPs. The risk-sensitive formulation includes a scalar parameter which we denote as γ. For small γ, to first order the risk-sensitive criterion is a weighted sum of the mean and variance of the cost. For $\gamma > 0$ the control will be "risk-averse", while for $\gamma < 0$ it will be "risk-seeking." As $\gamma \to 0$, we return

*Corresponding author Stefano P. Coraluppi. Tel. 781-273-3388; Fax 781-273-9345; E-mail stefano.coraluppi@alphatech.com.

[1]ALPHATECH Inc., 50 Mall Rd, Burlington, MA 02139, U.S.A.

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

to the risk-neutral formulation. Furthermore, it has been shown in [9] in the linear systems setting that there is a finite $\gamma_0 > 0$ such that the risk-sensitive controller approaches the optimal H_∞ controller as $\gamma \to \gamma_0$ from below. In [6] it is shown that, in the MDP setting, as $\gamma \to \infty$ the risk-sensitive controller approaches the minimax controller. (The connection between nonlinear risk-sensitive and robust control is explored in [12], [15], [16], [17]). Thus, one approach to balancing the objectives of minimizing expected cost incurred and worst-case cost incurred is to use a risk-sensitive controller with an appropriately selected risk-sensitive parameter.

In this paper we introduce a mixed risk-neutral/minimax control formulation for MDPs which we claim is more appealing than the risk-sensitive formulation as a criterion to achieve performance and robustness objectives. Our mixed risk-neutral/minimax formulation parallels the so-called mixed H_2/H_∞ criterion that has been introduced in the linear systems setting. See [21] and [7] for details.

The paper is organized as follows. In Section 2, we review the key results for risk-neutral and minimax control that will be useful to develop structural results and a solution methodology for the mixed risk-neutral/minimax problem. This problem is studied in Section 3. Section 4 addresses the effectiveness of the mixed risk-neutral/minimax and the risk-sensitive families of controllers to trade off the risk-neutral and minimax objectives. The paper ends with a brief conclusion.

2 Risk-Neutral 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 $C_{i,N} = \sum_{k=i}^{k=N-1} c_k(x_k, u_k) + c_N(x_N)$. On the infinite horizon, we will be considering a discounted cost structure with discount factor β, $0 < \beta < 1$. That is, $c_k(\cdot, \cdot) = \beta^k c(\cdot, \cdot)\ \forall k$. Throughout the paper, we will be interested in *minimizing* certain functions of the total cost of interest.

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 $k = 0$. A *policy* or *control law* is a sequence of (possibly non-deterministic) 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 [13], [4], and [18].

If the state is observed, that is $y_k = x_k$, there exists a Markov policy that is optimal, i.e. one that is independent of the past trajectory, given the state at time k. In the partially observed setting, the conditional distribution of the state given past observations is an information state. It is defined recursively as follows:

$$\pi_{k+1} = r(\pi_k, u_k, y_{k+1}) = \frac{\pi_k P(u_k) \bar{Q}(y_{k+1}, u_k)}{\pi_k P(u_k) \bar{Q}(y_{k+1}, u_k) \underline{1}} \tag{2}$$

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 an $|X|$-dimensional vector belonging to the space Π, the unit simplex in $\Re_+^{|X|}$, where $\Re_+^{|X|}$ is the set of $|X|$-tuples of non-negative real numbers. The value function, i.e. the "cost-to-go" associated with the objective (1) at time k, has the important properties that it is piecewise linear and concave in π_k. These properties can be exploited to construct an algorithm to determine the optimal policy. See [1], [19] and [14] for details.

The *minimax* objective is given by

$$\bar{J}(\mu,\pi_0) = \sup_{\omega\in\Omega_0^\mu} \sum_k c_k(x_k,u_k). \tag{3}$$

Here and below Ω_k^μ is the set of trajectories of the form $(x_k, u_k, x_{k+1}, u_{k+1}, \ldots)$ 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. Early work in minimax control includes [5], where the connection between stochastic and deterministic descriptions of uncertainty is explored. The partially observed problem can be addressed thanks to the results of [2]. This is pursued in [6], both for the finite and infinite horizon problems.

If the state of the system is observed, as with risk-neutral control there exists a Markov policy that is optimal. In the partially observed setting, an information state for the MDP with respect to (3) is given by the following recursion, where by $s[x]$ we mean the xth component of vector s:

$$s_0[x] = \begin{cases} 0 & \text{if } \pi_0[x] \neq 0, \\ -\infty & \text{otherwise,} \end{cases} \tag{4}$$

$$s_{k+1}[x'] = f_k(s_k, u_k, y_{k+1}). \tag{5}$$

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

$$f_k(s_k,u_k,y_{k+1}) = \begin{cases} \max_{x\in\tilde{X}(x',u_k)}[s_k[x] + c_k(x,u_k)] & \text{if } \tilde{X}(x',u_k) \neq \emptyset,\ x' \in \tilde{Y}(y_{k+1},u_k) \\ -\infty & \text{otherwise,} \end{cases} \tag{6}$$

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. The information state at each time k is a $|X|$-dimensional vector belonging to $\tilde{\Re}_+^{|X|}$, where $\tilde{\Re}_+ = \{\Re_+, -\infty\}$. The ith component can be interpreted as the worst-case cost incurred in reaching state i at time k, given all available information.

The minimax objective (3) depends on π_0 only to the extent that trajectories are feasible or not. Thus, by a slight abuse of notation, we prefer to denote the objective by $\bar{J}(\mu, s_0)$.

The *risk-sensitive* objective is given by

$$J^\gamma(\mu,\pi_0) = \frac{1}{\gamma}\log E^{\mu,\pi_0}[\exp{(\gamma \sum_k c_k(x,u))}]. \tag{7}$$

Later we shall examine properties of the risk-sensitive controller vis-a-vis the mixed risk-neutral/minimax controller, which we now introduce.

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 $\eta(s_0) \geq \eta_0(s_0)\ \forall s_0 \in \{0, -\infty\}^{|X|}$, where $\eta_0(s_0) = \min_{\mu \in M} \bar{J}(\mu, s_0)$. 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 $s_0 \in \{0, -\infty\}^{|X|}$,

$$\bar{J}(\mu, s_0) = \max_{\omega \in \Omega_0^\mu} \sum_k c_k(x, u)(\omega) \leq \eta(s_0). \tag{8}$$

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

$$J(\mu^\star, \pi_0) = \min_{\mu \in M(\eta(\cdot))} J(\mu, \pi_0), \tag{9}$$

for all $\pi_0 \in \Pi$. Again, recall that s_0 depends on π_0 as given by (4).

3.1 The Finite Horizon Case

In the general, partially observed setting, we wish now to address the task of determining an optimal policy $\mu^\star$ as defined by (9), 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 is our candidate information state (sufficient statistic). It is given by $\{g_k\}, k = 0, 1, \ldots$, where $g_k := (\pi_k, s_k)$.

We now introduce a number of definitions. We say that a policy is *separated* if it can be expressed as a sequence of mappings from the statistic defined above to the set U. Let $M(\eta(s_0), g, k) \subset M$ be the set of all policies such that

$$\max_{\omega \in \Omega_k^\mu} [s_k[x_k] + \sum_{l=k}^{N} c_l(x_l, u_l)](\omega) \leq \eta(s_0). \tag{10}$$

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 $M(\eta(s_0), g, k) \neq \emptyset$. 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:

$$V_{k,N}^{\eta(s_0)}(g) := \min_{\mu \in M(\eta, g, k)} E[\sum_{l=k}^{N} c_l(x_l, u_l) | g_k = g]. \tag{11}$$

In particular, we have

$$V_{0,N}^{\eta(s_0)}(g_0) = \min_{\mu \in M(\eta(s_0), g_0, 0)} J(\mu, \pi_0) = J(\mu^\star, \pi_0), \tag{12}$$

using equation (9).

Theorem 1 (Dynamic Programming). The value function defined in equation (11) satisfies the following dynamic programming equations for all feasible g:

$$
\begin{aligned}
V_{N,N}^{\eta(s_0)}(g) &= \pi \cdot c_N, & (13)\\
V_{k,N}^{\eta(s_0)}(g) &= \min_{u \in U(\eta(s_0),g,k)} E[c_k(x,u) + V_{k+1,N}^{\eta(s_0)}(g_{k+1})|g_k = g]. & (14)
\end{aligned}
$$

Furthermore, a policy $\mu_s^\star$ that achieves the minimum in equation (13) also achieves the minimum in (11). 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.

Proof. First we show that the value function satisfies (13), (14). We have

$$
\begin{aligned}
V_{k,N}^{\eta(s_0)}(g) &= \min_{\mu \in M(\eta(s_0),g,k)} E[c_k(x,u) + \sum_{l=k+1}^{N} c_l(x_l,u_l)|g_k = g]\\
&= \min_{\mu \in M(\eta(s_0),g,k)} E[c_k(x,u) + E[\sum_{l=k+1}^{N} c_l(x_l,u_l)|g_k = g, y_{k+1}]|g_k = g]\\
&= \min_{\mu \in M(\eta(s_0),g,k)} \{E[c_k(x,u)|g_k = g]\\
&\qquad +E[E[\sum_{l=k+1}^{N} c_l(x_l,u_l)|g_k = g, y_{k+1}]|g_k = g]\}\\
&= \min_{u \in U(\eta(s_0),g,k)} \{\pi \cdot c_k(u)\\
&\qquad +E[\min_{\mu \in M(\eta(s_0),g_{k+1},k+1)} E[\sum_{l=k+1}^{N} c_l(x_l,u_l)|g_{k+1}]|g_k = g]\}\\
&= \min_{u \in U(\eta(s_0),g,k)} \{\pi \cdot c_k(u) + E[V_{k+1,N}^{\eta(s_0)}(g_{k+1})|g_k = g]\}.
\end{aligned}
$$

Now assume that there is a separated policy $\mu_s^\star$ that satisfies (13), (14). It follows that $\mu_s^\star$ satisfies (11), and using (12) we see that $\mu_s^\star$ is an optimal policy within $M(\eta(s_0),(\pi_0,s_0),0)$. □

Note that for a given time k, the feasible information states g for which we are interested in the minimization in (14) 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 [1]. 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. 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

$$
f(x) = \| Ax \| \cdot g(\frac{Ax}{\| Ax \|}), \; x \in \Pi, \tag{15}
$$

is also concave. □

Using the Lemma, we can show the following.

Theorem 2 (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.

Proof. The proof is by induction. Note first that $V_{N,N}^{\eta(s_0)}(\pi, s) = \pi \cdot c_N$ is linear in π, so it is concave as well. Next, assume that $V_{k+1,N}^{\eta(s_0)}(\pi, s)$ is concave as a function of π, for $0 \leq k < N$. Using the Lemma and the fact that the sum of concave functions is concave, it follows that

$$E[c_k(x,u) + V_{k+1,N}^{\eta(s_0)}(f_k(\pi, u, y_{k+1}), s_{k+1})|g_k = g] \tag{16}$$

is concave as a function of π. Now by the Lemma and (14), it follows that $V_{k,N}^{\eta(s_0)}(\pi, s)$ is concave as a function of π. □

Theorem 3 (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.

Proof. Again the proof is by induction. $V_{N,N}^{\eta(s_0)}(\pi, s) = \pi \cdot c_N$ is clearly piecewise linear in π. Now, assume that $V_{k+1,N}^{\eta(s_0)}(\pi, s)$ is piecewise linear in π, for $0 \leq k < N$. Then we can express $V_{k+1,N}^{\eta(s_0)}(\pi, s)$ in the form

$$V_{k+1,N}^{\eta(s_0)}(\pi, s) = \min_{\alpha_{k+1} \in A_{k+1}(\eta_0(s_0), s)} \{\pi \cdot \alpha_k\}, \tag{17}$$

where $A_{k+1}(\eta(s_0), s)$ is a finite set of $|X|$-dimensional vectors. Using (2) and (14), we have

$$V_{k,N}^{\eta(s_0)}(s, \pi) = \min_{u \in U(\eta(s_0), g, k)} \sum_i \pi[i][c_k(i,u) + \sum_{y,j} P_{ij}(u) \bar{Q}_{jj}(y,u) \alpha_{k+1}^{l(\pi,u,y)}[j]], \tag{18}$$

where $\alpha_{k+1}^{l(\pi,u,y)} \in A_{k+1}(\eta(s_0), f_{k+1}(s, u, y))$ is the vector corresponding to the (risk-neutral) information state $r(\pi, u, y)$ at time $k+1$. We see that $V_{k,N}^{\eta(s_0)}$ is of the form (17), and is therefore piecewise linear. □

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 (13), (14). 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:

$$|s_k| = (2^{|X|} - 1)(|U| \cdot |Y|)^k. \tag{19}$$

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

$$|U|^{\frac{|Y|^{N-k}-1}{|Y|-1}}. \tag{20}$$

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

$$|A_k| \leq |U| \cdot |A_{k+1}|^{|Y|}. \tag{21}$$

Thus, the controller complexity at time k is bounded by the product of (19) and (20).

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 (14) 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 [10]). In the general, partially observed setting, it has been shown (see [6]) 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 (14):

$$V_k^{\eta(s_0)}(g) = \min_{u\in U(\eta(s_0),g,k)} E[c_k(x,u) + V_{k+1}^{\eta(s_0)}(g_{k+1})]. \tag{22}$$

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

$$V_{k+1}^{\eta(s_0)}(\pi, s) = \beta V_k^{\frac{\eta(s_0)}{\beta}}(\pi, \frac{s}{\beta}). \tag{23}$$

Combining (22) and (23), we obtain the following equation characterizing the value function:

$$V_k^{\eta(s_0)}(\pi, s) = \min_{u\in U(\eta(s_0),(s,\pi),0)} E[c(x,u) + \beta V_k^{\frac{\eta(s_0)}{\beta}}(r(\pi,u,y_{k+1}), \frac{f_k(s,u,y_{k+1})}{\beta})]. \tag{24}$$

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 (24) 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 4 (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 $\hat{\mu}$ satisfies the following, $\forall \pi_0 \in \Pi$:

$$J(\hat{\mu}, \pi_0) - J(\mu^\star, \pi_0) < \epsilon, \tag{25}$$

$$\bar{J}(\hat{\mu}, \pi_0) < \eta(s_0) + \epsilon, \tag{26}$$

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.

Proof. Choose $N > 0$ such that $\frac{\beta^N \|c\|}{1-\beta} < \epsilon$, where $\| c \|$ denotes the largest single stage cost, i.e. $\| c \| := \max_{x \in X, u \in U} |c(x,u)|$. Let $\hat{V}_0^{\eta(s_0)}(s_0, \pi_0)$ denote the value function associated with policy $\hat{\mu}$. Also, let $\tilde{V}_{0,N}^{\eta(s_0)}(s_0, \pi_0)$ denote the value function associated with executing policy $\hat{\mu}$ on $(0, N-1)$, with a terminal cost of $c_N = \frac{\beta^N \|c\|}{1-\beta}$. Then

$$\begin{aligned}
0 &\leq \hat{V}_0^{\eta(s_0)}(g_0) - V_0^{\eta(s_0)}(g_0) \\
&\leq \tilde{V}_{0,N}^{\eta(s_0)}(g_0) - V_0^{\eta(s_0)}(g_0) \\
&= V_{0,N}^{\eta(s_0)}(g_0) + \frac{\beta^N \| c \|}{1-\beta} - V_0^{\eta(s_0)}(g_0) \\
&\leq \frac{\beta^N \| c \|}{1-\beta} \\
&< \epsilon.
\end{aligned}$$

Next, note that the worst-case cost associated with policy $\hat{\mu}$ can be bounded as follows:

$$\begin{aligned}
\tilde{J}(\hat{\mu}, \pi_0) &= \max_{\omega \in \Omega_0^{\hat{\mu}}} \sum_k c_k(x,u)(\omega) \\
&\leq \max_{\omega \in \Omega_0^{\hat{\mu}}} \sum_{k=0}^{N-1} c_k(x,u)(\omega) + \frac{\beta^N \| c \|}{1-\beta} \\
&< \eta(s_0) + \epsilon.
\end{aligned}$$

Thus, given $\epsilon > 0$, we have constructed a near-optimal policy based on a finite horizon approximation, that is within ϵ of the performance of $\mu^\star$, and within ϵ of the worst-case bound $\eta(s_0)$. □

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.

4 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.

It is easy to see that the mixed risk-neutral/minimax controller has risk-neutral and minimax controllers as limiting cases. Specifically, as $\eta(s_0) \to \infty$ we have $M(\eta(s_0), g_0, 0) \to M$, so that $\lim_{\eta(s_0)\to\infty} V_{0,N}^{\eta(s_0)}(g_0) = V_{0,N}(\pi_0)$ using (12). That is, as we relax the constraint on worst-case behavior we recover the risk-neutral formulation. Similarly, as $\eta(s_0) \to \eta_0(s_0)$, the mixed risk-neutral/minimax controller will be an optimal minimax controller. Indeed, any policy satisfying (8) with bound $\eta_0(s_0)$, the lowest achievable minimax objective by any policy in M, must be a minimax policy. 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 $\gamma \to 0$ and $\gamma \to \infty$, respectively.

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 distribution $\pi_0 \in \Pi$ 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. It can be shown that if a family of policies is monotonic in a parameter, it is efficient.

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 (14) 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 $V_{k,N}^{\eta_2(s_0)}(g) \leq V_{k,N}^{\eta_1(s_0)}(g)$.

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

$$P(u) = \begin{bmatrix} 0.5 - \theta(u) & 2\theta(u) & 0.5 - \theta(u) \\ 0.5 - \theta(u) & 2\theta(u) & 0.5 - \theta(u) \\ 0.5 - \theta(u) & 2\theta(u) & 0.5 - \theta(u) \end{bmatrix}, \tag{27}$$

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.

Note that the family of mixed risk-neutral/minimax optimal policies is a family of deterministic policies, i.e. a deterministic mapping from available observations to the set of controls U. Since it is complete, it follows that every efficient policy is deterministic. Another simple way that this property can be established is the following.

Theorem 5 (Efficient Policies). Consider our standard MDP model, with the performance and robustness objectives as defined above. Every efficient policy is deterministic.

Proof. 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}, $\mu_{d2} \neq \mu_{d1}$, with $0 < p < 1$. 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,

$$r_{nd} = \max\{r_{d1}, r_{d2}\}. \tag{28}$$

The performance under policy μ_{nd} is given by

$$p_{nd} = p \cdot p_{d1} + (1-p) \cdot p_{d2}. \tag{29}$$

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. $\square$

5 Conclusion

This paper introduces the 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.

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

References

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

[2] 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.

[3] 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.

[4] D. P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 1995.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] M. R. James. Asymptotic analysis of nonlinear stochastic risk-sensitive control and differential games. *Mathematics of Control, Signals, and Systems*, 5:401–417, 1992.

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

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

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

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

[17] I. R. Petersen, M. R. James, and P. Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. In *Proc. IEEE Conference on Decision and Control*, 1997.

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

[19] 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.

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

[21] 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.