Robust Control of Markov Decision Processes and Connection to Risk-Sensitive Control

Stefano Coraluppi and Steven Marcus
Electrical Engineering Department and
Institute for Systems Research
University of Maryland
College Park, MD 20742
stefano@isr.umd.edu, marcus@isr.umd.edu

Abstract

This paper introduces a formulation of the robust control problem in the partially observed Markov Decision Process (POMDP) setting. We show that this formulation is the large risk limit of the risk-sensitive control problem. Exploiting this connection, we derive an information state process and dynamic programming equations for the value function. We develop a methodology to determine an optimal policy for finite horizon problems, and near-optimal policies on the infinite horizon. Finally, we introduce an alternative formulation of the robust control problem, leading to stationary policies on the infinite horizon, and provide a methodology to determine optimal policies in this setting.

1 Introduction

Robust control theory has been developed primarily in the linear systems context. It is essentially a minmax approach to control system design, whereby we choose a control to minimize the cost associated with worst-case disturbances. More recently, formulations in the nonlinear setting have been studied. A formulation for Finite State Machines has been proposed in [1]. In this paper, we propose a robust control formulation for MDPs which is precisely the minmax criterion studied in [2]. This formulation does not require much of the machinery in [1], and we believe it is a natural way to formulate the robust control problem in the MDP setting.

The connection between robust and risk-sensitive control in the linear systems setting was established in [6]. It is shown that there is a risk parameter value $\gamma_0 > 0$ such that the solution to the corresponding risk-sensitive control problem yields precisely the optimal H_∞ controller. We show that in the MDP context, the robust control problem is obtained in the limit as $\gamma \to \infty$ of the risk-sensitive problem.

In [1], an information state process and dynamic programming equations are derived for the partially observed, risk-sensitive control problem. We derive an information state process and dynamic programming equations for the robust control problem by taking their limiting form as the risk sensitive parameter tends to infinity. We develop a methodology to determine optimal policies which does not require structural results for the value function.

In general, in the infinite horizon setting there does not exist a stationary optimal policy for the robust control problem. Thus, we consider a finite horizon approximation with sufficiently large horizon, and prove that the resulting policy achieves near-optimality. Additionally, motivated by analogous work in risk-sensitive control (see [8], [3], [7], [4]), we introduce an alternative formulation of the robust control problem for which we show that, on the infinite horizon, there exists a stationary optimal policy. We introduce solution methodologies for this formulation.

2 Problem Formulation

We are interested in discrete-time stochastic dynamical systems with a finite state space and control space. This class of systems can be described by a controlled Markov chain. This is given by $M = (X, Y, U, \{P(u), u \in U\}, \{Q(u), u \in U\})$, where X is the finite state space, Y is the finite output space, and U is the finite set of controls. $P(u)$ and $Q(u)$ are the state transition matrix and the output matrix, respectively, for $u \in U$. That is, $P_{ij}(u) = pr(x_{t+1} = j | x_t = i, u_t = u)$ and $Q_{ij}(u) = pr(y_t = j | x_t = i, u_{t-1} = u)$.

We denote by π_0 the initial distribution of x_0 and define the random variable $C_{k,N} = \sum_{t=k}^{N-1} c_t(x_t, u_t)$. A policy g is a sequence of mappings $g = (g_0, g_1, \ldots)$ such that $u_k = g_k(y^k), y^k = (y_1, \ldots, y_k)$. We denote by G the set of all policies. Given a policy g and an initial distribution π_0, let $p^g_{\pi_0}(\cdot)$ be the corresponding probability distribution on trajectories of the system.

Our objective is to determine the optimal policy or control law with respect to the following cost function:

$$\tilde{J}(g, \pi_0) = \max_{\omega \in \Omega, p^g_{\pi_0}(\omega) \neq 0} C_{0,N}(\omega), \tag{1}$$

where Ω is the set of all trajectories for the system. The maximization is over all possible values that the random variable $C_{0,N}$ takes with nonzero probability. That is, we seek the control law which minimizes the worst case cost incurred.

3 Risk-Sensitive Control

In order to address the task of determining the optimal policy with respect to (1), we will exploit a number of results from risk-sensitive control theory, which we briefly summarize in this section.

Consider the following risk-sensitive cost function:

$$J^\gamma(g,\pi_0) = \frac{1}{\gamma}\log E^g_{\pi_0}[\exp(\gamma C_{0,N})] \tag{2}$$

This cost function is a generalization of the classical stochastic control criterion, which is obtained in the limit $\gamma \to 0$ and which is commonly referred to as a risk-neutral cost. We will consider the following equivalent criterion for $\gamma > 0$:

$$\hat{J}^\gamma(g,\pi_0) = E^g_{\pi_0}[\exp(\gamma C_{0,N})], \gamma > 0. \tag{3}$$

In the fully observed setting, there exists a Markov policy g that is optimal. If the state of the system is partially observed, then we reformulate the problem in terms of an information state. Let $p = |Y|$. A reference measure is defined in [1], under which all observations $y \in Y$ are independent and equiprobable at every time. With this reference measure, which we denote by $\dagger$, we can express the total cost as follows:

$$\begin{aligned}\hat{J}^\gamma(g,\pi_0) &= E^g_{\pi_0}[\exp(\gamma C_{0,N})] \\ &= E^\dagger_{\pi_0}[\lambda_N \exp(\gamma C_{0,N})],\end{aligned} \tag{4}$$

where $\lambda_N := p^N \prod_{t=1}^N Q_{x_t y_t}(u_{t-1})$. Define

$$\sigma_k(i) := E^\dagger_{\pi_0}[\mathbf{1}[x_k = i]\lambda_k \exp(\gamma C_{0,k})|y_1,\ldots,y_k]. \tag{5}$$

It is shown in [1] that $\sigma_k, k \in \{0,\ldots\}$ constitutes an information state process. It evolves in time as follows:

$$\sigma_0 = \pi_0, \tag{6}$$

$$\sigma_{k+1} = p\sigma_k D^\gamma(k,u_k)\bar{Q}(y_{k+1},u_k), \tag{7}$$

where $\bar{Q}$ is a diagonal matrix with $\bar{Q}_{ii}(y,u_k) = pr(y_{k+1} = y|x_{k+1} = i, u_k = u)$. There exists a separated policy that is optimal, and the value function is piecewise linear and convex in the information state (see [5]). See [4] for an algorithm to determine optimal policies in this setting.

4 Robust Control

Our first step is to establish that the large risk limit of the risk sensitive control problem is precisely the robust control problem. For this, we need a modified version of the Varadhan-Laplace lemma, which is a standard result in analysis which we state without proof.

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

$$\lim_{\gamma\to\infty}\frac{1}{\gamma}\log\sum_{\omega\in\Omega}\exp[\gamma F^\gamma(\omega)] = \max_{\omega\in\Omega} F(\omega) \tag{8}$$

Lemma 2 (Modified Varadhan-Laplace). 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 $\forall \omega \in \Omega$, independent of γ. Then

$$\lim_{\gamma\to\infty} \frac{1}{\gamma} \log \sum_{\omega\in\Omega} p(\omega) \exp\left[\gamma F^\gamma(\omega)\right] = \max_{\omega\in\Omega, p(\omega)\neq 0} F(\omega) \tag{9}$$

Proof

$$\begin{aligned}
\lim_{\gamma\to\infty} \frac{1}{\gamma} \log \sum_{\omega\in\Omega} p(\omega) \exp\left[\gamma F^\gamma(\omega)\right] &= \lim_{\gamma\to\infty} \frac{1}{\gamma} \log \sum_{\omega\in\Omega, p(\omega)\neq 0} p(\omega) \exp\left[\gamma F^\gamma(\omega)\right] \\
&= \lim_{\gamma\to\infty} \frac{1}{\gamma} \log \sum_{\omega\in\Omega, p(\omega)\neq 0} \exp\left(\gamma[F^\gamma(\omega) + \frac{\log p(\omega)}{\gamma}]\right) \\
&= \max_{\omega\in\Omega, p(\omega)\neq 0} F(\omega)
\end{aligned}$$

by Lemma 1. □

Remark In particular, if $\sum_{\omega\in\Omega} p(\omega) = 1$, i.e. $p(\cdot)$ is a probability measure on Ω, it follows that

$$\lim_{\gamma\to\infty} \frac{1}{\gamma} \log E[\exp\left(\gamma F^\gamma(\omega)\right)] = \max_{\omega\in\Omega, p(\omega)\neq 0} F(\omega) \tag{10}$$

The following theorem establishes the connection between the risk-sensitive control and robust control of MDPs.

Theorem 1 (Cost Function). Let $J^\infty(\cdot,\cdot) = \lim_{\gamma\to\infty} J^\gamma(\cdot,\cdot)$. Let $\tilde{J}(\cdot,\cdot)$ be given by (1). Then $J^\infty = \tilde{J}$.

Proof. Let a policy g and an initial distribution on the states π_0 be given. Let Ω denote the set of possible evolutions of the system, and let $p^g(\omega)$ denote the probability under policy g of the evolution $\omega \in \Omega$.

$$\begin{aligned}
J^\infty(g, \pi_0) &= \lim_{\gamma\to\infty} \frac{1}{\gamma} \log E^g_{\pi_0}[\exp(\gamma C_{0,N})] \\
&= \max_{\omega\in\Omega, p^g_{\pi_0}(\omega)\neq 0} C_{0,N}(\omega)
\end{aligned}$$

using Lemma 2. □

Remark. Note that the dependence of the cost on the state transition matrix and the initial distribution π_0 is limited to whether starting locations and trajectories are feasible or infeasible. That is, for each evolution of the system, its probability $p^g_{\pi_0}(\omega)$ of occuring is relevant only to the extent that it is zero or nonzero.

We are now in a position to derive an information state process and dynamic programming equations for the robust control problem. First, we derive a statistic by taking the large risk limit of a function of the risk-sensitive information state.

The statistic will turn out to be an information state for the MDP with respect to (1).

Theorem 2 (Statistic). Let $\tilde{X}(x', u, k)$ be the set of states at time k from which, using control $u \in U$, there is a nonzero probability that the next state of the system will be x'. Also, let $\tilde{Y}(y, u, k)$ denote the set of states at time k that can result in observation y at time k, if the previous control at time $k-1$ is u.

A statistic for the MDP is given by $\{s_k\},\ k = 0, 1, \ldots$, where $\forall x,\ x' \in X$,

$$s_0(x) = \begin{cases} 0 & \text{if } \pi_0(x) \neq 0 \\ -\infty & \text{otherwise} \end{cases} \tag{11}$$

$$s_{k+1}(x') = \begin{cases} \max_{x \in \tilde{X}(x',u,k)} [s_k(x) + c_k(x,u)] & \text{if } \tilde{X}(x',u,k) \neq \emptyset \text{ and } x' \in \tilde{Y}(y,u,k+1) \\ -\infty & \text{otherwise} \end{cases} \tag{12}$$

Proof. Let

$$s_k = \lim_{\gamma \to \infty} \frac{1}{\gamma} \log \sigma_k \tag{13}$$

and apply Varadhan-Laplace. Details are omitted for lack of space.

Remark. At time k, if a particular state i is feasible, $s_k(i)$ is the worst case cost incurred in the system's evolution from a feasible starting state to state i at time k. If the state is infeasible, the worst case cost incurred is $-\infty$.

The following result relates the information state and total cost.

Theorem 3 (Cost, Information State).

$$\tilde{J}(g, \pi_0) = \max_{y^N} \max_{i \in X} s_N(i) \tag{14}$$

Proof.

$$\begin{aligned} \tilde{J}(g, \pi_0) &= \lim_{\gamma \to \infty} \frac{1}{\gamma} \log E^{\dagger}_{\pi_0} [\exp(\gamma C_{0,N})] \\ &= \lim_{\gamma \to \infty} \frac{1}{\gamma} \log E^{\dagger}_{\pi_0} [\sum_{i \in X} \sigma_N(i)] \\ &= \lim_{\gamma \to \infty} \frac{1}{\gamma} \log \sum_{y^N} p^N \sum_{i \in X} \sigma_N(i) \\ &= \lim_{\gamma \to \infty} \frac{1}{\gamma} \log \sum_{y^N} \sum_{i \in X} p^N \exp \gamma s_N(i) \\ &= \max_{y^N} \max_{i \in X} s_N(i) \quad \square \end{aligned}$$

We define the value function as

$$Z_{k,N}(s) := \min_{g \in G} \max_{y_{k+1}, \ldots, y_N} \max_{i \in X} s_N(i), \text{ where } s_k = s \tag{15}$$

Thus we have

$$Z_{0,N}(s_0) = Z_{0,N}(\lim_{\gamma\to\infty} \frac{1}{\gamma}\log\pi_0) = \min_{g\in G} \tilde{J}(g,\pi_0) \tag{16}$$

Again by a limiting procedure, we can obtain dynamic programming equations characterizing the value function for the system. First we introduce the following notation for the set of all information states. Define $\tilde{R}_+^{|X|} := \{R_+, -\infty\}^{|X|}$, where R_+ is the set of nonnegative reals.

Theorem 4 (Dynamic Programming). The value function satisfies the following, $\forall s \in \tilde{R}_+^{|X|}$:

$$Z_{N,N}(s) = \max_{i\in X} s(i) \tag{17}$$

$$Z_{k,N}(s) = \min_{u\in U} \max_{y\in Y} Z_{k+1,N}(s_{k+1}(s,u,y)) \tag{18}$$

where $s_{k+1}(s,u,y)$ is given by (12). Furthermore, a policy that achieves the minimum in equations (17) and (18), also achieves the minimum in (15). That is, an optimal separated policy is optimal within the larger class G of all policies.

Proof. Omitted for lack of space.

The fully observed case. In the special case when the state of the system is observed, equations (11),(12) take the following form:

$$s_0(x) = \begin{cases} 0 & \text{if } x = x_0 \\ -\infty & \text{otherwise} \end{cases} \tag{19}$$

$$s_{k+1}(x') = \begin{cases} \max_{x\in\tilde{X}(x',u,k)}[s_k(x) + c_k(x,u)] & \text{if } x' = x_{k+1} \\ -\infty & \text{otherwise} \end{cases} \tag{20}$$

The total cost can be expressed in terms of the information state as follows:

$$\tilde{J}(g,\pi_0) = \max_{i\in X} s_N(i) \tag{21}$$

The value function takes the form

$$Z_{k,N}(s) = \min_{g\in G} \max_{i\in X} s_N(i), \text{ where } s_k = s, \tag{22}$$

with dynamic programming equations given by:

$$Z_{N,N}(s) = \max_{i\in X} s(i) \tag{23}$$

$$Z_{k,N}(s) = \min_{u\in U} Z_{k+1,N}(s_{k+1}(s,u)) \tag{24}$$

The minimization in (24) is unaffected by scaling the information state s. If we normalize the information state, we see that the optimality of a separated policy reduces

to the optimality of a Markov policy, as one would expect. Note that we could have obtained the same results for this special case by directly utilizing the risk-sensitive, fully observed formulation.

Remark. The results in the fully observed case are not new. See [2] for early work on the problem.

For MDPs on an infinite horizon, we consider a formulation with discounted costs. We have

$$\tilde{J}(g, \pi_0) = \sup_{\omega \in \Omega, p^g_{\pi_0}(\omega) \neq 0} C_{0,\infty}(\omega), \tag{25}$$

with $C_{0,\infty}$ given by $C_{0,\infty} = \sum_{k=0}^{\infty} \beta^k c(x_k, u_k), 0 < \beta < 1$. Theorems 1, 2 are valid on the infinite horizon as well. We define the infinite horizon value function as follows:

$$Z_k(s) := \lim_{N \to \infty} Z_{k,N}(s) \tag{26}$$

We can verify that the limit in (26) is well-defined by noting that $\lim_{N\to\infty} S_{k,N}$ is well defined, and $Z_{k,N} = \lim_{\gamma \to \infty} \frac{1}{\gamma} \log S^{\gamma}_{k,N}(\exp(\gamma s))$. Thus

$$Z_k(s) = \lim_{\gamma \to \infty} \frac{1}{\gamma} \log S^{\gamma}_k(\exp(\gamma s)) \tag{27}$$

We can relate the value function to the cost function (25) by taking the limit in (16) as $N \to \infty$. We obtain:

$$Z_0(s_0) = \inf_{g \in G} \tilde{J}(g, \pi_0) \tag{28}$$

The following result characterizes the infinite horizon value function.

Theorem 5. The value function satisfies the following, $\forall s \in \tilde{R}_+^{|X|}$:

$$Z_0(s) = \min_{u \in U} \max_{y \in Y} \beta Z_0\left(\frac{s_1(s, u, y)}{\beta}\right) \tag{29}$$

Proof. Omitted for lack of space.

5 Optimal Robust Policies

In risk-neutral and risk-sensitive control, the determination of optimal policies for partially observed MDPs requires the use of structural results for the value function. Without such results, the minimization over a continuum of information states (the unit simplex), is intractable.

In the robust 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

$$|s_k| = (2^{|X|} - 1)(|U|\ |Y|)^k \tag{30}$$

feasible states. Note that this number grows exponentially in the size of the horizon.

A simple 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 (17) and (18) to find the optimal control at each state of interest.

In the special case when the state is fully observed, we can solve (23) and (24) and find a Markov policy that is optimal. Of course, in this case the number of states of interest at each time k remains constant irrespective of the horizon size.

Let us now consider the determination of optimal robust policies on the infinite horizon. Unfortunately, it turns out (as in risk-sensitive control) that in general the optimal policy is nonstationary. In the risk-sensitive setting, this can be understood as a progressive discounting of the risk- sensitive parameter as time increases. In the robust setting, discounted costs affect the optimal choice of control as well.

Our approach to determining an optimal infinite horizon policy will be to approximate the problem by a finite horizon one with sufficiently large horizon, and solving the corresponding optimal control problem. The following result establishes the fact that this methodology gives a policy with a cost that is arbitrarily close to optimal.

Theorem 6. Consider the MDP given by $M = (X, Y, U, \{P(u), u \in U\}, \{Q(u), u \in U\})$, with cost function (25). Let $\epsilon > 0$ be given, and set $N = \max\{\lceil \xi \rceil, 1\}$, where $\xi = \frac{\log[\frac{(1-\beta)\epsilon}{\|c\|}]}{\log \beta}$. Let $\hat{g}$ denote the policy that achieves the minimum in eqns. (17), (18), and is arbitrary but fixed on $(N, \ldots)$. Let $\hat{Z}_k, k = 0, 1, \ldots$ denote the associated value function. Then we have, $\forall s \in \tilde{R}_+^{|X|}$,

$$0 \leq \hat{Z}_0(s) - Z_0(s) \leq \epsilon. \tag{31}$$

Proof. Omitted for lack of space.

6 An Alternate Robust Control Formulation

On the infinite horizon, with discounted costs, our robust control formulation leads to a nonstationary optimal policy in general. This is a somewhat surprising result, and it is shared by the risk-sensitive formulation as well. In the risk-sensitive setting, an interpretation for this is that the discounting of costs amounts to a discounting of the

risk-sensitive parameter. As time increases, the control problem reverts to the risk-neutral one. A formulation which leads to a stationary optimal policy while retaining elements of risk-sensitivity is discussed in [8]. An extension of the formulation in the partially observed setting is introduced in [4].

Motivated by the work in risk-sensitive control, we would like to introduce a formulation for the robust control problem which will lead to a stationary infinite horizon policy, while still preserving aspects of robustness. Specifically, we set the cost-to-go from a state i at time k to be the cost incurred at time k, plus the discounted worst case future cost. In the fully observed setting, this takes the following form:

$$\tilde{h}_{k,N}(i) = \min_{u \in U}\{c_k(i,u) + \beta \max_{j \in \tilde{X}'(i,u)} \tilde{h}_{k+1,N}(j)\} \tag{32}$$

$$\tilde{h}_{N,N}(i) = c_N(i) \tag{33}$$

where $\tilde{X}'(i,u)$ 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. We now show that this formulation is the large risk limit of the risk-sensitive alternate formulation.

Theorem 7. Define $h^{\infty}_{k,N} := \lim_{\gamma \to \infty} h^{\gamma}_{k,N}$, where $h^{\gamma}_{k,N}$ is given by the following:

$$h^{\gamma}_{k,N}(i) = \min_{u \in U}\{c_k(i,u) + \frac{\beta}{\gamma} \log \sum_{j \in X} P_{ij}(u) \exp\left(\gamma h^{\gamma}_{k+1,N}(j)\right)\} \tag{34}$$

$$h^{\gamma}_{N,N}(i) = c_N(i) \tag{35}$$

Then $\tilde{h}_{k,N}(i) = h^{\infty}_{k,N}(i),\ \forall i \in X, 0 \leq k \leq N$.

Proof. Follows by invoking the Modified Varadhan-Laplace Lemma.

In a similar way as in the risk-sensitive formulation, it can be shown that for the infinite horizon with stationary costs, we have $\lim_{N \to \infty} \tilde{h}_{k,N} = \tilde{h}$ independent of k, so that an optimal policy satisfies the following equation, $\forall i \in X$:

$$\tilde{h}(i) = \min_{u \in U}\{c(i,u) + \beta \max_{j \in \tilde{X}'(i,u)} \tilde{h}(j)\} \tag{36}$$

Furthermore, there exists an optimal stationary policy that can be determined through value iteration. At each step of value iteration, the evaluation of the cost associated with a policy can be determined by using the contractive mapping $T^g[z](\cdot)$, where for each $i \in X$,

$$T^g[z](i) = \{c(i,g(i)) + \beta \max_{j \in \tilde{X}'(i,g(i))} z(j)\} \tag{37}$$

In the partially observed setting, this formulation takes the following form, $\forall \pi \in \Pi$, where Π is the $|X|$-dimensional unit simplex.

$$\tilde{h}_{k,N}(\pi) = \min_{u \in U}\{\pi \cdot c_k(u) + \beta \max_{\pi' \in \tilde{X}'(\pi,u)} \tilde{h}_{k+1,N}(\pi')\} \tag{38}$$

$$\tilde{h}_{N,N}(\pi) = \pi \cdot c_N \tag{39}$$

where $\tilde{X}'(\pi, u)$ is the set of information states that the system reaches in one transition with nonzero probability, given that the probability distribution on the current states is given by π, and control u is used.

On the infinite horizon with stationary costs, the partially observed formulation takes the following form:

$$\tilde{h}(\pi) = \min_{u \in U} \{\pi \cdot c(u) + \beta \max_{\pi' \in \tilde{X}'(\pi,u)} \tilde{h}(\pi')\} \tag{40}$$

Structural results are not yet available for the partially observed value function. Thus, both on the finite and infinite horizon, an approximate optimal policy must be determined numerically.

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

References

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

[2] 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.

[3] 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.

[4] S. P. Coraluppi and S. I. Marcus. Risk-sensitive control of markov decision processes. In *Proc. 30th Conference on Information Sciences and Systems*, 1996.

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

[6] 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.

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

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