CORC Tech Report TR-2002-07
Robust dynamic programming[*]

G. Iyengar [†]

Submitted Dec. 3rd, 2002. Revised May 4, 2004.

Abstract

In this paper we propose a robust formulation for discrete time dynamic programming (DP). The objective of the robust formulation is to systematically mitigate the sensitivity of the DP optimal policy to ambiguity in the underlying transition probabilities. The ambiguity is modeled by associating a set of conditional measures with each state-action pair. Consequently, in the robust formulation each policy has a set of measures associated with it. We prove that when this set of measures has a certain "Rectangularity" property all the main results for finite and infinite horizon DP extend to natural robust counterparts. We identify families of sets of conditional measures for which the computational complexity of solving the robust DP is only modestly larger than solving the DP, typically logarithmic in the size of the state space. These families of sets are constructed from the confidence regions associated with density estimation, and therefore, can be chosen to guarantee any desired level of confidence in the robust optimal policy. Moreover, the sets can be easily parameterized from historical data. We contrast the performance of robust and non-robust DP on small numerical examples.

1 Introduction

This paper is concerned with sequential decision making in uncertain environments. Decisions are made in stages and each decision, in addition to providing an immediate reward, changes the context of future decisions; thereby affecting the future rewards. Due to the uncertain nature of the environment, there is limited information about both the immediate reward from each decision and the resulting future state. In order to achieve a good performance over all the stages the decision maker has to trade-off the immediate payoff with future payoffs. Dynamic programming (DP) is the mathematical framework that allows the decision maker to efficiently compute a good overall strategy by succinctly encoding the evolving information state. In the DP formalism the uncertainty in the environment is modeled by a Markov process whose transition probability depends both on the information state and the action taken by the decision maker. It is assumed that the transition probability corresponding to each state-action pair is known to the decision maker, and the goal is to choose a policy, i.e. a rule that maps states to actions, that maximizes some performance measure. Puterman (1994) provides a excellent introduction to the DP formalism and its various applications. In this paper, we assume that the reader has some prior knowledge of DP.

[*] Submitted to *Math. Oper. Res.* . Do not distribute

[†] IEOR Department, Columbia University, Email: garud@ieor.columbia.edu. Research partially supported by NSF grants CCR-00-09972 and DMS-01-04282.

The DP formalism encodes information in the form of a "reward-to-go" function (see Puterman, 1994, for details) and chooses an action that maximizes the sum of the immediate reward and the expected "reward-to-go". Thus, to compute the optimal action in any given state the "reward-to-go" function for all the future states must be known. In many applications of DP, the number of states and actions available in each state are large; consequently, the computational effort required to compute the optimal policy for a DP can be overwhelming – Bellman's "curse of dimensionality". For this reason, considerable recent research effort has focused on developing algorithms that compute an approximately optimal policy efficiently (Bertsekas and Tsitsiklis, 1996; de Farias and Van Roy, 2002).

Fortunately, for many applications the DP optimal policy can be computed with a modest computational effort. In this paper we restrict attention to this class of DPs. Typically, the transition probability of the underlying Markov process is estimated from historical data and is, therefore, subject to statistical errors. In current practice, these errors are ignored and the optimal policy is computed assuming that the estimate is, indeed, the true transition probability. The DP optimal policy is quite sensitive to perturbations in the transition probability and ignoring the estimation errors can lead to serious degradation in performance (Nilim and El Ghaoui, 2002; Tsitsiklis et al., 2002). Degradation in performance due to estimation errors in parameters has also been observed in other contexts (Ben-Tal and Nemirovski, 1997; Goldfarb and Iyengar, 2003). Therefore, there is a need to develop DP models that explicitly account for the effect of errors.

In order to mitigate the effect of estimation errors we assume that the transition probability corresponding to a state-action pair is not exactly known. The ambiguity in the transition probability is modeled by associating a set $\mathcal{P}(s, a)$ of conditional measures with each state-action pair (s, a). (We adopt the convention of the decision analysis literature wherein *uncertainty* refers to random quantities with *known* probability measures and *ambiguity* refers to unknown probability measures (see, e.g. Epstein and Schneider, 2001)). Consequently, in our formulation each policy has a set of measures associated with it. The value of a policy is the minimum expected reward over the set of associated measures, and the goal of the decision maker is to choose a policy with maximum value, i.e. we adopt a maximin approach. We will refer to this formulation as *robust* DP. We prove that, when the set of measures associated with a policy satisfy a certain "Rectangularity" property (Epstein and Schneider, 2001), the following results extend to natural robust counterparts: the Bellman recursion, the optimality of deterministic policies, the contraction property of the value iteration operator, and the policy iteration algorithm. "Rectangularity" is a sort of independence assumption and is a minimal requirement for these results to hold. However, this assumption is not always appropriate, and is particularly troublesome in the infinite horizon setting (see Appendix A for details). We show that if the decision maker is restricted to stationary policies the effects of the "Rectangularity" assumption are not serious.

There is some previous work on modeling ambiguity in the transition probability and mitigating its effect on the optimal policy. Satia and Lave (1973); White and Eldieb (1994); Bagnell et al. (2001) investigate ambiguity in the context of infinite horizon DP with finite state and action spaces. They model ambiguity by constraining the transition probability matrix to lie in a pre-specified polytope. They do not discuss how one constructs this polytope. Moreover, the complexity of the resulting robust DP is at least an order of magnitude higher than DP. Shapiro and Kleywegt (2002) investigate ambiguity in the context of stochastic programming and propose a sampling based method for solving the maximin problem. However, they do not discuss how to choose and calibrate the set of ambiguous priors. None of this work discusses the dynamic structure of the ambiguity; in particular, there is no discussion of the central role of "Rectangularity". Our theoretical contributions are based on recent work on uncertain priors in the economics literature (Gilboa and Schmeidler, 1989; Epstein and Schneider, 2001, 2002; Hansen and Sargent, 2001). The focus of this body of work is on the axiomatic justification for uncertain priors in the context of multi-period utility maximization. It does not provide any means of selecting the set of uncertain priors nor does it focus on efficiently solving the resulting robust DP.

In this paper we identify families of sets of conditional measures that have the following desirable properties. These families of sets provide a means for setting any desired level of confidence in the robust optimal policy. For a given confidence level, the corresponding set from each family is easily parameterizable from data. The complexity of solving the robust DP corresponding to these families of sets is only modestly larger that the non-robust counterpart. These families of sets are constructed from the confidence regions associated with density estimation.

While this paper was being prepared for publication we became aware of a technical report by Nilim and El Ghaoui (2002) where they formulate finite horizon robust DP in the context of an aircraft routing problem. A "robust counterpart" for the Bellman equation appears in their paper but they do not justify that this "robust counterpart", indeed, characterizes the robust value function. Like all the previous work on robust DP, Nilim and El Ghaoui also do not recognize the importance of Rectangularity. However, they do introduce sets based on confidence regions and show that the finite horizon robust DP corresponding to these sets can be solved efficiently.

The paper has two distinct and fairly independent parts. The first part comprising of Section 2 and Section 3 presents the robust DP theory. In Section 2 we formulate finite horizon robust DP and the "Rectangularity" property that leads to the robust counterpart of the Bellman recursion; and Section 3 formulates the robust extension of discounted infinite horizon DP. The focus of the second part comprising of Section 4 and Section 5 is on computation. In Section 4 we describe three families of sets of conditional measures that are based on the confidence regions, and show that the computational effort required to solve the robust DP corresponding to these sets is only modestly higher than that required to solve the non-robust counterpart. The results in this section, although independently obtained, are not new and were first obtained by Nilim and El Ghaoui (2002). In Section 5 we provide basic examples and computational results. Section 6 includes some concluding remarks.

2 Finite horizon robust dynamic programming

Decisions are made at discrete points in time $t \in T = \{0, 1, \ldots\}$ referred to as decision epochs. In this section we assume that T finite, i.e. $T = \{0, \ldots, N-1\}$ for some $N \geq 1$. At each epoch $t \in T$ the system occupies a state $s \in \mathcal{S}_t$, where $\mathcal{S}_t$ is assumed to be discrete (finite or countably infinite). In a state $s \in \mathcal{S}_t$ the decision maker is allowed to choose an action $a \in \mathcal{A}_t(s)$, where $\mathcal{A}_t(s)$ is assumed to be discrete. Although many results in this paper extend to non-discrete state and action sets, we avoid this generality because the associated measurability issues would detract from the ideas that we want to present in this work.

For any discrete set $\mathcal{B}$, we will denote the set of probability measures on $\mathcal{B}$ by $\mathcal{M}(\mathcal{B})$. Decision makers can choose actions either randomly or deterministically. A random action is a state $s \in \mathcal{S}_t$ corresponds to an element $q_s \in \mathcal{M}(\mathcal{A}(s))$ with the interpretation that an action $a \in \mathcal{A}(s)$ is selected with probability $q_s(a)$. Degenerate probability measures that assign all the probability mass to a single action correspond to deterministic actions.

Associated with each epoch $t \in T$ and state-action pair (s, a), $a \in \mathcal{A}(s)$, $s \in \mathcal{S}_t$, is a set of conditional measures $\mathcal{P}_t(s, a) \subseteq \mathcal{M}(\mathcal{S}_{t+1})$ with the interpretation that if at epoch t, action a is chosen in state s, the state s_{t+1} at the next epoch $t+1$ is determined by some conditional measure $p_{sa} \in \mathcal{P}_t(s, a)$. Thus, the state transition is *ambiguous*. (We adopt the convention of the decision analysis literature wherein *uncertainty* refers to random quantities with *known* probability measures and *ambiguity* refers to unknown probability measures (see, e.g. Epstein and Schneider, 2001)).

The decision maker receives a reward $r_t(s_t, a_t, s_{t+1})$ when the action $a_t \in \mathcal{A}(s_t)$ is chosen in state $s_t \in \mathcal{S}$ at the decision epoch t, and the state at the next epoch is $s_{t+1} \in \mathcal{S}$. Since s_{t+1} is ambiguous, we allow the reward at time t to depend on s_{t+1} as well. Note that one can assume, without loss of generality, that the reward $r_t(\cdot, \cdot, \cdot)$ is certain. The reward $r_N(s)$ at the epoch N is a only a function of the state $s \in \mathcal{S}_N$.

We will refer to the collection of objects $\big\{T, \{\mathcal{S}_t, \mathcal{A}_t, \mathcal{P}_t, r_t(\cdot, \cdot, \cdot) : t \in T\}\big\}$ as a finite horizon *ambiguous Markov decision process* (AMDP). The notation above is a modification of that in Puterman (1994) and the structure of ambiguity is motivated by Epstein and Schneider (2001).

A decision rule d_t is a procedure for selecting actions in each state at a specified decision epoch $t \in T$. We will call a decision rule history dependent if it depends on the entire past history of the system as represented by the sequence of past states and actions, i.e. d_t is a function of the history $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$. Let $\mathcal{H}_t$ denote the set of all histories h_t. Then a randomized decision rule d_t is a map $d_t : \mathcal{H}_t \mapsto \mathcal{M}(\mathcal{A}(s_t))$. A decision rule d_t is called deterministic if it puts all the probability mass on a single action $a \in \mathcal{A}(s_t)$, and Markovian if it is a function of the current state s_t alone.

The set of all conditional measures consistent with a deterministic Markov decision rule d_t is given by

$$\mathcal{T}^{d_t} = \Big\{\mathbf{p} : \mathcal{S}_t \mapsto \mathcal{M}(\mathcal{S}_{t+1}) : \forall s \in \mathcal{S}_t,\ \mathbf{p}_s \in \mathcal{P}_t(s, d_t(s))\Big\}, \tag{1}$$

i.e. for every state $s \in \mathcal{S}$, the next state can be determined by any $p \in \mathcal{P}_t(s, d_t(s))$. The set of all conditional measures consistent with a history dependent decision rule d_t is given by

$$\mathcal{T}^{d_t} = \left\{\mathbf{p} : \mathcal{H}_t \mapsto \mathcal{M}(\mathcal{A}(s_t) \times \mathcal{S}_{t+1}) : \begin{array}{ll} \forall h \in \mathcal{H}_t, & \mathbf{p}_h(a, s) = q_{d_t(h)}(a) p_{s_t a}(s), \\ & p_{s_t a} \in \mathcal{P}(s_t, a),\ a \in \mathcal{A}(s_t),\ s \in \mathcal{S}_{t+1} \end{array} \right\} \tag{2}$$

A policy prescribes the decision rule to be used at all decision epochs. Thus, a policy π is a sequence of decision rules, i.e. $\pi = (d_t : t \in T)$. Given the ambiguity in the conditional measures, a policy π induces a collection of measure on the history space $\mathcal{H}_N$. We assume that the set $\mathcal{T}^\pi$ of measures consistent with a policy π has the following structure.

Assumption 1 (Rectangularity) *The set $\mathcal{T}^\pi$ of measures consistent with a policy π is given by*

$$\begin{aligned} \mathcal{T}^\pi &= \Big\{\mathbf{P} : \forall h_N \in \mathcal{H}_N,\ \mathbf{P}(h_N) = \prod_{t \in T} \mathbf{p}_{h_t}(a_t, s_{t+1}),\ \mathbf{p}_{h_t} \in \mathcal{T}^{d_t},\ t \in T\Big\}, \\ &= \mathcal{T}^{d_0} \times \mathcal{T}^{d_1} \times \cdots \times \mathcal{T}^{d_{N-1}}, \end{aligned} \tag{3}$$

where the notation in (3) simply denotes that each $p \in \mathcal{T}^\pi$ is a product of $p_t \in \mathcal{T}^{d_t}$, and vice versa.

The Rectangularity assumption is motivated by the structure of the recursive multiple priors in Epstein and Schneider (2001). We will defer discussing the implications of the this assumption until after we define the objective of the decision maker.

The reward $V_0^\pi(s)$ generated by a policy π starting from the initial state $s_0 = s$ is defined as follows.

$$V_0^\pi(s) = \inf_{\mathbf{P} \in \mathcal{T}^\pi} \mathbf{E}^{\mathbf{P}}\Big[\sum_{t \in T} r_t(s_t, d_t(h_t), s_{t+1}) + r_N(s_N)\Big], \tag{4}$$

where $\mathbf{E}^{\mathbf{P}}$ denotes the expectation with respect to the fixed measure $\mathbf{P} \in \mathcal{T}^\pi$. Equation (4) defines the reward of a policy π to be the minimum expected reward over all measures consistent with the policy π. Thus, we take a worst-case approach in defining the reward. In the optimization literature this approach is known as the *robust* approach (Ben-Tal and Nemirovski, 1998). Let Π denote the set of all history dependent policies. Then the goal of *robust* DP is to characterize the *robust* value function

$$V_0^*(s) = \sup_{\pi\in\Pi}\Big\{V_0^\pi(s)\Big\} = \sup_{\pi\in\Pi}\left\{\inf_{\mathbf{P}\in\mathcal{T}^\pi}\mathbf{E}^{\mathbf{P}}\left[\sum_{t\in T} r_t(s_t, d_t(h_t), s_{t+1}) + r_N(s_N)\right]\right\}, \tag{5}$$

and an optimal policy π^* if the supremum is achieved.

In order to appreciate the implications of the Rectangularity assumption the objective (5) has to interpreted in an adversarial setting: the decision maker chooses π; an adversary observes π, and chooses a measure $\mathbf{P}\in\mathcal{T}^\pi$ that minimizes the reward. In this context, Rectangularity is a form of an independence assumption: the choice of particular distribution $\bar{p}\in\mathcal{P}(s_t, a_t)$ in a state-action pair (s_t, a_t) at time t does not limit the choices of the adversary in the future. This, in turn, leads to a separability property that is crucial for establishing the robust counterpart of the Bellman recursion (see Theorem 1). Such a model for an adversary is not always appropriate. See Appendix A for an example of such a situation. We will return to this issue in the context of infinite horizon models in Section 3.

The *optimistic* value $\bar{V}_0^\pi(s_0)$ of a policy π starting from the initial state $s_0 = s$ is defined as

$$\bar{V}_0^\pi(s) = \sup_{\mathbf{P}\in\mathcal{T}^\pi}\mathbf{E}^{\mathbf{P}}\left[\sum_{t\in T} r_t(s_t, d_t(h_t), s_{t+1}) + r_N(s_N)\right]. \tag{6}$$

Let $V_0^\pi(s_0;\mathbf{P})$ denote the non-robust value of a policy π corresponding to a particular choice $\mathbf{P}\in\mathcal{T}^\pi$. Then $\bar{V}_0^\pi(s_0) \geq V_0^\pi(s_0;\mathbf{P}) \geq V_0^\pi(s_0)$. Analogous to the robust value function $V_0^*(s)$, the optimistic value function $\bar{V}_0^*(s)$ is defined as

$$\bar{V}_0^*(s) = \sup_{\pi\in\Pi}\Big\{\bar{V}_0^\pi(s)\Big\} = \sup_{\pi\in\Pi}\left\{\sup_{\mathbf{P}\in\mathcal{T}^\pi}\mathbf{E}^{\mathbf{P}}\left[\sum_{t\in T} r_t(s_t, d_t(h_t), s_{t+1}) + r_N(s_N)\right]\right\}. \tag{7}$$

Remark 1 *Since our interest is in computing the robust optimal policy π^*, we will restrict attention to the robust value function V_0^*. However, all the results in this paper imply a corresponding result for the optimistic value function $\bar{V}_0^*$ with the* $\inf_{\mathbf{P}\in\mathcal{T}^\pi}(\cdot)$ *replaced by* $\sup_{\mathbf{P}\in\mathcal{T}^\pi}(\cdot)$.

Let $V_n^\pi(h_n)$ denote the reward obtained by using policy π over epochs $n, n+1, \ldots, N-1$, starting from the history h_n, i.e.

$$V_n^\pi(h_n) \quad = \quad \inf_{\mathbf{P}\in\mathcal{T}_n^\pi}\mathbf{E}^{\mathbf{P}}\left[\sum_{t=n}^{N-1} r_t(s_t, d_t(h_t), s_{t+1}) + r_N(s_N)\right], \tag{8}$$

where Rectangularity implies that the set of conditional measures $\mathcal{T}_n^\pi$ consistent with the policy π and the history h_n is given by

$$\begin{aligned}\mathcal{T}_n^\pi \quad &= \quad \left\{\mathbf{P}_n : \mathcal{H}_n \mapsto \prod_{t=n}^{N-1}(\mathcal{A}_t\times\mathcal{S}_{t+1}) : \begin{array}{ll}\forall h_n\in\mathcal{H}_n, & \mathbf{P}_{h_n}(a_n, s_{n+1}, \ldots, a_{N-1}, s_N) = \prod_{t=n}^{N-1}\mathbf{p}_{h_t}(a_t, s_{t+1}), \\ & \mathbf{p}_{h_t}\in\mathcal{T}^{d_t}, t = n, \ldots, N-1\end{array}\right\}, \\ &= \quad \mathcal{T}^{d_n}\times\mathcal{T}^{d_1}\times\cdots\times\mathcal{T}^{d_{N-1}}, \\ &= \quad \mathcal{T}^{d_n}\times\mathcal{T}_{n+1}^\pi.\end{aligned} \tag{9}$$

Let $V_n^*(h_n)$ denote the optimal reward starting from the history h_n at the epoch n, i.e.

$$V_n^*(h_n) = \sup_{\pi\in\Pi_n}\Big\{V_n^\pi(h_n)\Big\} = \sup_{\pi\in\Pi_n}\left\{\inf_{\mathbf{P}\in\mathcal{T}_n^\pi}\mathbf{E}^{\mathbf{P}}\left[\sum_{t=n}^{N-1} r_t(s_t, d_t(h_t), s_{t+1}) + r_N(s_N)\right]\right\}, \tag{10}$$

where Π_n is the set of all history dependent randomized policies for epochs $t\geq n$.

Theorem 1 (Bellman equation) *The set of functions $\{V_n^* : n = 0, 1, \ldots, N\}$ satisfies the following robust Bellman equation:*

$$\begin{aligned} V_N^*(h_N) &= r_N(s_N), \\ V_n^*(h_n) &= \sup_{a \in \mathcal{A}(s_n)} \left\{ \inf_{p \in \mathcal{P}(s_n, a)} \mathbf{E}^p \Big[r_n(s_n, a, s) + V_{n+1}^*(h_n, a, s) \Big] \right\}, \quad n = 0, \ldots, N-1. \end{aligned} \tag{11}$$

Proof: From (9) it follows that

$$V_n^*(h_n) = \sup_{\pi \in \Pi} \left\{ \inf_{\mathbf{P} = (\mathbf{p}, \bar{\mathbf{P}}) \in \mathcal{T}^{d_n} \times \mathcal{T}_{n+1}^{\pi}} \mathbf{E}^{\mathbf{P}} \Big[\sum_{t=n}^{N-1} r_t(s_t, d_t(h_t), s_{t+1}) + r_N(s_N) \Big] \right\}.$$

Since the conditional measures $\bar{\mathbf{P}}$ do not affect the first term $r_n(s_n, d_n(h_n), s_{n+1})$, we have:

$$\begin{aligned} V_n^*(h_n) &= \sup_{\pi \in \Pi_n} \left\{ \inf_{(\mathbf{p}, \bar{\mathbf{P}}) \in \mathcal{T}^{d_n} \times \mathcal{T}_{n+1}^{\pi}} \mathbf{E}^{\mathbf{p}} \Big[r_n(s_n, d_n(h_n), s_{n+1}) + \mathbf{E}^{\bar{\mathbf{P}}} \Big[\sum_{t=n+1}^{N-1} r_t(s_t, d_t(h_t), s_{t+1}) + r_N(s_N) \Big] \Big] \right\}, \\ &= \sup_{\pi \in \Pi_n} \left\{ \inf_{\mathbf{p} \in \mathcal{T}^{d_n}} \mathbf{E}^{\mathbf{p}} \Big[r_n(s_n, d_n(h_n), s_{n+1}) + \inf_{\bar{\mathbf{P}} \in \mathcal{T}_{n+1}^{\pi}} \mathbf{E}^{\bar{\mathbf{P}}} \Big[\sum_{t=n+1}^{N-1} r_t(s_t, d_t(h_t), s_{t+1}) + r_N(s_N) \Big] \Big] \right\}, \\ &= \sup_{\pi \in \Pi_n} \left\{ \inf_{\mathbf{p} \in \mathcal{T}^{d_n}} \mathbf{E}^{\mathbf{p}} \Big[r_n(s_n, d_n(h_n), s_{n+1}) + V_{n+1}^{\pi}(h_n, d_n(h_n), s_{n+1}) \Big] \right\}, \end{aligned} \tag{12}$$

where the last equality follows from the definition of $V_{n+1}^{\pi}(h_{n+1})$ in (8).

Let $(d_n(h_n)(\omega), s_{n+1}(\omega))$ denote any realization of the random action-state pair corresponding the (randomized) decision rule d_n. Then $V_{n+1}^{\pi}(h_n, d_n(h_n)(\omega), s_{n+1}(\omega)) \leq V_{n+1}^*(h_n, d_n(h_n)(\omega), s_{n+1}(\omega))$. Therefore, (12) implies that

$$\begin{aligned} V_n^*(h_n) &\leq \sup_{\pi \in \Pi_n} \left\{ \inf_{\mathbf{p} \in \mathcal{T}^{d_n}} \mathbf{E}^{\mathbf{p}} \Big[r_n(s_n, d_n(h_n), s_{n+1}) + V_{n+1}^*(h_n, d_n(h_n), s_{n+1}) \Big], \right. \\ &= \sup_{d_n \in \mathcal{D}_n} \left\{ \inf_{\mathbf{p} \in \mathcal{T}^{d_n}} \mathbf{E}^{\mathbf{p}} \Big[r_n(s_n, d_n(h_n), s_{n+1}) + V_{n+1}^*(h_n, d_n(h_n), s_{n+1}) \Big] \right\}, \end{aligned} \tag{13}$$

where $\mathcal{D}_n$ is the set of all history dependent decision rules at time n, and (13) follows from the fact that the term within the expectation only depends on $d_n \in \mathcal{D}_n$.

Since $V_{n+1}^*(h_{n+1}) = \sup_{\pi \in \Pi_{n+1}} \left\{ V_{n+1}^{\pi}(h_{n+1}) \right\}$, it follows that for all $\epsilon > 0$ there exists a policy $\pi_{n+1}^{\epsilon} \in \Pi_{n+1}$ such that $V_{n+1}^{\pi_{n+1}^{\epsilon}}(h_{n+1}) \geq V_{n+1}^*(h_{n+1}) - \epsilon$, for all $h_{n+1} \in \mathcal{H}_{n+1}$. For all $d_n \in \mathcal{D}_n$, $(d_n, \pi_{n+1}^{\epsilon}) \in \Pi_n$. Therefore,

$$\begin{aligned} V_n^*(h_n) &= \sup_{\pi \in \Pi_n} \left\{ \inf_{\mathbf{p} \in \mathcal{T}^{d_n}} \mathbf{E}^{\mathbf{p}} \Big[r_n(s_n, d_n(h_n), s_{n+1}) + V_{n+1}^{\pi}(h_n, d_n(h_n), s_{n+1}) \Big], \right. \\ &\geq \sup_{d_n \in \mathcal{D}_n} \left\{ \inf_{\mathbf{p} \in \mathcal{T}^{d_n}} \mathbf{E}^{\mathbf{p}} \Big[r_n(s_n, d_n(h_n), s_{n+1}) + V_{n+1}^{\pi_{n+1}^{\epsilon}}(h_n, d_n(h_n), s_{n+1}) \Big] \right\}, \\ &\geq \sup_{d_n \in D_n} \left\{ \inf_{\mathbf{p} \in \mathcal{T}^{d_n}} \mathbf{E}^{\mathbf{p}} \Big[r_n(s_n, d_n(h_n), s_{n+1}) + V_{n+1}^*(h_n, d_n(h_n), s_{n+1}) \Big] \right\} - \epsilon. \end{aligned} \tag{14}$$

Since $\epsilon > 0$ is arbitrary, (13) and (14) imply that

$$V_n^*(h_n) = \sup_{d_n \in \mathcal{D}_n} \left\{ \inf_{\mathbf{p} \in \mathcal{T}^{d_n}} \mathbf{E}^{\mathbf{p}} \Big[r_n(s_n, d_n(h_n), s_{n+1}) + V_{n+1}^*(h_n, d_n(h_n), s_{n+1}) \Big] \right\}.$$

The definition of $\mathcal{T}^{d_n}$ in (2) implies that $V_n^*(h_n)$ can be rewritten as follows.

$$\begin{aligned} V_n^*(h_n) &= \sup_{q\in\mathcal{M}(\mathcal{A}(s_n))} \inf_{p_{s_n a}\in\mathcal{P}_n(s_n,a)} \Big\{ \sum_{a\in\mathcal{A}(s_n)} q(a) \Big[\sum_{s\in\mathcal{S}} p_{s_n a}(s)\big[r_n(s_n,a,s) + V^*_{n+1}(h_n,a,s)\big]\Big]\Big\}, \\ &= \sup_{q\in\mathcal{M}(\mathcal{A}(s_n))} \Big\{ \sum_{a\in\mathcal{A}(s_n)} q(a) \inf_{p_{s_n a}\in\mathcal{P}_n(s_n,a)} \Big[\sum_{s\in\mathcal{S}} p_{s_n a}(s)\big[r_n(s_n,a,s) + V^*_{n+1}(h_n,a,s)\big]\Big]\Big\}, \\ &= \sup_{a\in\mathcal{A}(s_n))} \Big\{ \inf_{p\in\mathcal{P}_n(s_n,a)} \Big[\sum_{s\in\mathcal{S}} p(s)\big[r_n(s_n,a,s) + V^*_{n+1}(h_n,a,s)\big]\Big]\Big\}, \qquad (15) \end{aligned}$$

where (15) follows from the fact that

$$\sup_{u\in W} w(u) \geq \sum_{u\in W} q(u)w(u),$$

for all discrete sets W, functions $w : W \mapsto \mathbf{R}$, and probability measures q on W. ■

While this paper was being prepared for publication we became aware of a technical report by Nilim and El Ghaoui (2002) where they formulate robust solutions to finite-horizon AMDPs with finite state and action spaces. A "robust counterpart" of the Bellman equation appears in their paper. This "robust counterpart" reduces to the robust Bellman equation (11) provided one assumes that the set of measures $\mathcal{P}(s,a)$ is convex. The convexity assumption is very restrictive, e.g. a discrete set of measures $\mathcal{P}(s,a) = \{q_1,\ldots,q_m\}$ is not convex. Moreover, they do not prove that the solution $V_t(s)$ of the "robust counterpart" is the robust value function, i.e. there exists a policy that achieves $V_t(s)$. Their paper does not discuss the dynamic structure of the ambiguity; in particular, there is no discussion of the structure of the set $\mathcal{T}^\pi$ of measures consistent with a policy. The robust Bellman equation characterizes the robust value function if and only if $\mathcal{T}^\pi$ satisfies Rectangularity, it would be impossible to claim that the solution of a recursion is the robust value function without invoking Rectangularity is some form. In summary, while the robust solutions to AMDPs were addressed in Nilim and El Ghaoui (2002), we provide the necessary theoretical justification for the robust Bellman recursion and generalize the result to countably infinite state and action sets.

The following corollary establishes that one can restrict the decision maker to deterministic policies without affecting the achievable robust reward.

Corollary 1 *Let Π_D be the set of all history dependent deterministic policies. Then Π_D is adequate for characterizing the value function V_n in the sense that for all $n = 0,\ldots,N-1$,*

$$V_n^*(h_n) = \sup_{\pi\in\Pi_D} \Big\{ V_n^\pi(h_n)\Big\}.$$

Proof: This result follows from (11). The details are left to the reader. ■

Next, we show that it suffices to restrict oneself to deterministic Markov policies, i.e. policies where the deterministic decision rule d_t at any epoch t is a function of only the current state s_t.

Theorem 2 (Markov optimality) *For all $n = 0,\ldots,N$, the robust value function $V_n^*(h_n)$ is a function of the current state s_n alone, and $V_n^*(s_n) = \sup_{\pi\in\Pi_{MD}} \{V_n^\pi(s_n)\}$, $n \in T$, where Π_{MD} is the set of all deterministic Markov policies. Therefore, the robust Bellman equation (11) reduces to*

$$V_n^*(s_n) = \sup_{a\in\mathcal{A}(s_n)} \Big\{ \inf_{p\in\mathcal{P}_n(s_n,a)} \mathbf{E}^p \Big[r_n(s_n,a,s) + V^*_{n+1}(s)\Big]\Big\}, \quad n \in T. \qquad (16)$$

Proof: The result is established by induction on the epoch t. For $t = N$, the value function $V_N^*(h_N) = r_N(s_N)$ and is, therefore, a function of only the current state.

Next, suppose the result holds for all $t > n$. From the Bellman equation (11) we have

$$\begin{aligned} V_n^*(h_n) &= \sup_{a \in \mathcal{A}(s_n)} \left\{ \inf_{p \in \mathcal{P}_n(s_n,a)} \mathbf{E}^p \Big[r_n(s_n,a,s) + V_{n+1}^*(h_n,a,s) \Big] \right\}, \\ &= \sup_{a \in \mathcal{A}(s_n)} \left\{ \inf_{p \in \mathcal{P}_n(s_n,a)} \mathbf{E}^p \Big[r_n(s_n,a,s) + V_{n+1}^*(s) \Big] \right\}, \qquad (17) \end{aligned}$$

where (17) follows from the induction hypothesis. Since the right hand side of (17) depends on h_n only via s_n, the result follows. ■

The recursion relation (16) forms the basis for robust DP. This relation establishes that, provided $V_{n+1}^*(s')$ is known for all $s' \in \mathcal{S}$, computing $V_n^*(s)$ reduces to a collection of optimization problems. Suppose the action set $\mathcal{A}(s)$ is finite. Then the optimal decision rule d_n^* at epoch n is given by

$$d_n^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}(s)} \left\{ \inf_{p \in \mathcal{P}_n(s,a)} \mathbf{E}^p \Big[r_n(s,a,s') + V_{n+1}(s') \Big] \right\}.$$

Hence, in order to compute the value function V_n^* efficiently one must be able to efficiently solve the optimization problem $\inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p[v]$ for a specified $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$ and $v \in \mathbf{R}^{|\mathcal{S}|}$. In Section 4 we describe three families of sets $\mathcal{P}(s,a)$ of conditional measures for which $\inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p[v]$ can be solved efficiently.

As noted in Remark 1, Theorem 2 implies the following result for the optimistic value function $\bar{V}_n^*$.

Theorem 3 *For $n = 0, \ldots, N$, the optimistic value function $\bar{V}_n^*(h_n)$ is a function of the current state s_n alone, and*

$$\bar{V}_n^*(s_n) = \sup_{\pi \in \Pi_{MD}} \left\{ \bar{V}_n^\pi(s_n) \right\}, \quad n \in T,$$

where Π_{MD} is the set of all deterministic Markov policies. Therefore,

$$\bar{V}_n^*(s_n) = \sup_{a \in \mathcal{A}(s_n)} \left\{ \sup_{p \in \mathcal{P}_n(s_n,a)} \mathbf{E}^p \Big[r_n(s_n,a,s) + \bar{V}_{n+1}^*(s) \Big] \right\}, \quad n \in T. \qquad (18)$$

3 Infinite horizon robust dynamic programming

In this section we formulate robust infinite horizon robust DP with a discounted reward criterion and describe methods for solving this problem. Robust infinite horizon DP with finite state and action spaces was addressed in Satia (1968); Satia and Lave (1973). A special case of the robust DP where the decision maker is restricted to stationary policies appears in Bagnell et al. (2001). We will contrast our contributions with the previous work as we establish the main results of this section.

The setup is similar to the one introduced in Section 2. As before, we assume that the decisions epochs are discrete, however now the set $T = \{0, 1, 2, \ldots, \} = \mathbf{Z}_+$. The system state $s \in \mathcal{S}$, where $\mathcal{S}$ is assumed to be discrete, and in state $s \in \mathcal{S}$ the decision maker is allowed to take a randomized action chosen from a discrete set $\mathcal{A}(s)$. As the notation suggests, in this section we assume that the state space is not a function of the decision epoch $t \in T$.

Unlike in the finite horizon setting, we assume that the set of conditional measures $\mathcal{P}(s,a) \subseteq \mathcal{M}(\mathcal{S})$ is not a function of the decision epoch $t \in T$. We continue to assume that the set $\mathcal{T}^\pi$ of measures consistent with a policy π satisfies Rectangularity, i.e. $\mathcal{T}^\pi = \prod_{t \in T} \mathcal{T}^{d_t}$. Note that Rectangularity implies that the adversary is allowed to choose a possibly different conditional measure $p \in \mathcal{P}(s,a)$ every time the state-action pair (s,a) is encountered. Hence we will refer to this adversary model as the *dynamic* model. In many applications of robust DP the transition probability is, in fact, fixed but the decision maker is only able to estimate to within a set. In such situations the dynamic model is not appropriate (see Appendix A for a discussion). Instead, one would prefer a *static* model where the adversary is restricted to choose the same, but unknown, $p_{sa} \in \mathcal{P}(s,a)$ every time the state-action pair (s,a) is encountered. We contrast the implications of the two models in Lemma 3. Bagnell et al. (2001) also has some discussion on this issue.

As before, the reward $r(s_t, a_t, s_{t+1})$ is a function of the current state s_t, the action $a_t \in \mathcal{A}(s_t)$, and the future state s_{t+1}; however, it is not a function of the decision epoch t. We will also assume that the reward is bounded, i.e. $\sup_{s,s' \in \mathcal{S}, a \in \mathcal{A}(s)}\{r(s,a,s')\} = R < \infty$. The reward $V^\pi(s)$ received by employing a policy π when the initial state $s_0 = s$ is given by

$$V_\lambda^\pi(s) = \inf_{\mathbf{P}\in\mathcal{T}^\pi} \mathbf{E}^{\mathbf{P}}\Big[\sum_{t=0}^{\infty} \lambda^t r(s_t, d_t(h_t), s_{t+1})\Big], \tag{19}$$

where $\lambda \in (0,1)$ is the discount factor. It is clear that for all policies π, $\sup_{s\in\mathcal{S}}\{V_\lambda^\pi(s)\} \leq \frac{R}{1-\lambda}$. The optimal reward in state s is given by

$$V_\lambda^*(s) = \sup_{\pi\in\Pi}\Big\{V^\pi(s)\Big\} = \sup_{\pi\in\Pi}\Big\{\inf_{\mathbf{P}\in\mathcal{T}^\pi} \mathbf{E}^{\mathbf{P}}\Big[\sum_{t=0}^{\infty} \lambda^t r(s_t, d_t(h_t), s_{t+1})\Big]\Big\}, \tag{20}$$

where Π is the set of all history dependent randomized policies. The optimistic value function $\bar{V}_\lambda^*$ can be defined as follows.

$$\bar{V}_\lambda^*(s) = \sup_{\pi\in\Pi}\Big\{\sup_{\mathbf{P}\in\mathcal{T}^\pi} \mathbf{E}^{\mathbf{P}}\Big[\sum_{t=0}^{\infty} \lambda^t r(s_t, d_t(h_t), s_{t+1})\Big]\Big\}. \tag{21}$$

As noted in Remark 1, all the results in this section imply a corresponding result for the optimistic value function $\bar{V}_\lambda^*$ with the $\inf_{\mathbf{P}\in\mathcal{T}^\pi}(\cdot)$ replaced by $\sup_{\mathbf{P}\in\mathcal{T}^\pi}(\cdot)$.

The following result is the infinite horizon counterpart of Theorem 2.

Theorem 4 (Markov optimality) *The decision maker can be restricted to deterministic Markov policies without any loss in performance, i.e.* $V_\lambda^*(s) = \sup_{\pi\in\Pi_{MD}}\{V_\lambda^\pi(s)\}$, *where* Π_{MD} *is the set of all deterministic Markov policies.*

Proof: Since $\mathcal{P}(s,a)$ only depends on the current state-action pair, this result follows from robust extensions of Theorem 5.5.1, Theorem 5.5.3 and Proposition 6.2.1 in Puterman (1994). ■

Let $\mathbf{V}$ denote the set of all bounded real valued functions on the discrete set $\mathcal{S}$. Let $\|V\|$ denote the L_∞ norm on $\mathbf{V}$, i.e.

$$\|V\| = \max_{s\in\mathcal{S}} |V(s)|.$$

Then $(\mathbf{V}, \|\cdot\|)$ is a Banach space. Let $\mathcal{D}$ be any subset of all deterministic Markov decision rules. Define the robust Bellman operator $\mathcal{L}_\mathcal{D}$ on $\mathbf{V}$ as follows: For all $V \in \mathbf{V}$,

$$\mathcal{L}_\mathcal{D} V(s) = \sup_{d\in\mathcal{D}}\Big\{\inf_{p\in\mathcal{P}(s,d(s))} \mathbf{E}^p\big[r(s,d(s),s') + \lambda V(s')\big]\Big\}, \quad s \in \mathcal{S}. \tag{22}$$

Theorem 5 (Bellman equation) *The operator* $\mathcal{L}_\mathcal{D}$ *satisfies the following properties:*

(a) The operator $\mathcal{L}_\mathcal{D}$ *is contraction mapping on* $\mathbf{V}$*; in particular, for all* $U, V \in \mathbf{V}$,

$$\|\mathcal{L}U - \mathcal{L}V\| \leq \lambda\|U - V\|. \tag{23}$$

(b) The operator equation $\mathcal{L}_\mathcal{D} V = V$ *has a unique solution. Moreover,*

$$V(s) = \sup_{\{\pi : d_t^\pi \in \mathcal{D}\}} \inf_{\mathbf{P}\in\mathcal{T}^\pi} \mathbf{E}^{\mathbf{P}}\Big[\sum_{t=0}^{\infty} \lambda^t r(s_t, d_t(h_t), s_{t+1})\Big],$$

where $\mathcal{T}^\pi$ *is defined in (3).*

Proof: Let $U, V \in \mathbf{V}$. Fix $s \in \mathcal{S}$, and assume that $\mathcal{L}U(s) \geq \mathcal{L}V(s)$. Fix $\epsilon > 0$ and choose $d \in \mathcal{D}$ such that for all $s \in \mathcal{S}$,

$$\inf_{p \in \mathcal{P}(s,d(s))} \mathbf{E}^{p}\big[r(s,d(s),s') + \lambda U(s')\big] \geq \mathcal{L}_{\mathcal{D}}U(s) - \epsilon.$$

Choose a conditional probability measure $p_s \in \mathcal{P}(s, d(s))$, $s \in \mathcal{S}$, such that

$$\mathbf{E}^{p_s}\big[r(s,d(s),s') + \lambda V(s')\big] \leq \inf_{p \in \mathcal{P}(s,d(s))} \mathbf{E}^{p}\big[r(s,d(s),s') + \lambda V(s')\big] + \epsilon.$$

Then

$$\begin{aligned}
0 \leq \mathcal{L}U(s) - \mathcal{L}V(s) &\leq \Big(\inf_{p \in \mathcal{P}(s,d(s))} \mathbf{E}^{p}\big[r(s,d(s),s') + \lambda U(s')\big] + \epsilon\Big) - \Big(\inf_{p \in \mathcal{P}(s,d(s))} \mathbf{E}^{p}\big[r(s,d(s),s') + \lambda V(s')\big]\Big),\\
&\leq \Big(\mathbf{E}^{p_s}\big[r(s,d(s),s') + \lambda U(s')\big] + \epsilon\Big) - \Big(\mathbf{E}^{p_s}\big[r(s,d(s),s') + \lambda V(s')\big] - \epsilon\Big),\\
&= \lambda \mathbf{E}^{p_s}[U - V] + 2\epsilon,\\
&\leq \lambda \mathbf{E}^{p_s}|U - V| + 2\epsilon,\\
&\leq \lambda \|U - V\| + 2\epsilon.
\end{aligned}$$

Repeating the argument for the case $\mathcal{L}U(s) \leq \mathcal{L}V(s)$ implies that

$$|\mathcal{L}U(s) - \mathcal{L}V(s)| \leq \lambda\|U - V\| + 2\epsilon, \quad \forall s \in \mathcal{S},$$

i.e. $\|\mathcal{L}U - \mathcal{L}V\| \leq \lambda\|U - V\| + 2\epsilon$. Since ϵ was arbitrary, this establishes part (a) of the Theorem.

Since $\mathcal{L}_{\mathcal{D}}$ is a contraction operator on a Banach space, the Banach fixed point theorem implies that the operator equation $\mathcal{L}_{\mathcal{D}}V = V$ has a unique solution $V \in \mathbf{V}$.

Fix π such that $d_t^\pi \in \mathcal{D}$, for all $t \geq 0$. Then

$$\begin{aligned}
V(s) &= \mathcal{L}_{\mathcal{D}}V(s),\\
&\geq \inf_{p_0 \in \mathcal{P}(s,d_0^\pi(s))} \mathbf{E}^{p_0}\big[r(s,d_0^\pi(s),s_1) + \lambda V(s_1)\big], &(24)\\
&\geq \inf_{p_0 \in \mathcal{P}(s,d_0^\pi(s))} \mathbf{E}^{p_0}\Big[r(s,d_0^\pi(s),s_1) + \lambda \inf_{p_1 \in \mathcal{P}(s_1,d_1^\pi(s_1))} \mathbf{E}^{p_1}\big[r(s_1,d_1^\pi(s_1),s_2) + \lambda V(s_2)\big]\Big], &(25)\\
&= \inf_{\mathbf{P} \in \mathcal{T}^\pi} \mathbf{E}^{\mathbf{P}}\Big[\sum_{t=0}^{1} r(s_t,d_t^\pi(s_t),s_{t+1}) + \lambda^2 V(s_{t+1})\Big], &(26)
\end{aligned}$$

where (24) follows from the fact that choosing a particular action $d_0^\pi(s)$ can only lower the value of the right hand side, (25) follows by iterating the same argument once more, and (26) follows from the Rectangularity assumption. Thus, for all $n \geq 0$,

$$\begin{aligned}
V(s) &\geq \inf_{\mathbf{P} \in \mathcal{T}^\pi} \mathbf{E}^{\mathbf{P}}\Big[\sum_{t=0}^{n} r(s_t,d_t^\pi(s_t),s_{t+1}) + \lambda^{n+1} V(s_{t+1})\Big],\\
&= \inf_{\mathbf{P} \in \mathcal{T}^\pi} \mathbf{E}^{\mathbf{P}}\Big[\sum_{t=0}^{\infty} r(s_t,d_t^\pi(s_t),s_{t+1}) + \lambda^{n+1} V(s_{t+1}) - \sum_{t=n+1}^{\infty} \lambda^t r(s_t,d_t^\pi(s_t),s_{t+1})\Big],\\
&\geq V^\pi(s) - \lambda^{n+1}\|V\| - \frac{\lambda^{n+1}R}{1-\lambda},
\end{aligned}$$

where $R = \sup_{s,s' \in \mathcal{S}, a \in \mathcal{A}(s)}\{r(s,a,s')\} < \infty$. Since n is arbitrary, it follows that

$$V(s) \geq \sup_{\{\pi : d_t^\pi \in \mathcal{D}, \forall t\}} \Big\{V^\pi(s)\Big\}. \tag{27}$$

```
The Robust Value Iteration Algorithm:
Input: $V \in \mathbf{V}$, $\epsilon > 0$
Output: $\tilde{V}$ such that $\|\tilde{V} - V^*\| \leq \frac{\epsilon}{2}$
 For each $s \in \mathcal{S}$, set $\tilde{V}(s) = \sup_{a \in \mathcal{A}(s)} \Big\{ \inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p\big[r(r,a,s') + \lambda V(s')\big]\Big\}$.
 while $\Big(\|\tilde{V} - V\| \geq \frac{(1-\lambda)}{4\lambda} \cdot \epsilon\Big)$ do
         $V = \tilde{V}$
         $\forall s \in \mathcal{S}$, set $\tilde{V}(s) = \sup_{a \in \mathcal{A}(s)} \Big\{ \inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p\big[r(r,a,s') + \lambda V(s')\big]\Big\}$.
 end while
 return $\tilde{V}$
```

Figure 1: Robust value iteration algorithm

Fix $\epsilon > 0$ and choose a deterministic decision rule $d \in \mathcal{D}$ such that for all $s \in \mathcal{S}$

$$V(s) = \mathcal{L}_{\mathcal{D}} V(s) \leq \inf_{p \in \mathcal{P}(s,d(s))} \mathbf{E}^p\Big[r(s,d(s),s') + \lambda V(s')\Big] + \epsilon.$$

Consider the policy $\pi = (d, d, \ldots)$. An argument similar to the one above establishes that for all $n \geq 0$

$$V(s) \leq V^{\pi}(s) + \lambda^n \|V\| + \frac{\epsilon}{1-\lambda}. \tag{28}$$

Since ϵ and n are arbitrary, it follows from (27) and (28) that $V(s) = \sup_{\{\pi : d_t^{\pi} \in \mathcal{D}, \forall t\}} \big\{V^{\pi}(s)\big\}$. ■

Corollary 2 *The properties of the operator $\mathcal{L}_{\mathcal{D}}$ imply the following:*

(a) Let d be any deterministic decision rule. Then the value V_{λ}^{π} of the stationary policy $\pi = (d, d, \ldots)$ is the unique solution of the operator equation

$$V(s) = \inf_{p \in \mathcal{P}(s,d(s))} \mathbf{E}^p\big[r(s,d(s),s') + \lambda V(s')\big], \quad s \in S. \tag{29}$$

(b) The value function V_{λ}^{*} is the unique solution of the operator equation

$$V(s) = \sup_{a \in \mathcal{A}(s)} \inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p\big[r(s,a,s') + \lambda V(s')\big], \quad s \in S. \tag{30}$$

Moreover, for all $\epsilon > 0$, there exists an ϵ-optimal stationary policy, i.e. there exists $\pi^{\epsilon} = (d^{\epsilon}, d^{\epsilon}, \ldots)$ such that $V_{\lambda}^{\pi^{\epsilon}} \geq V_{\lambda}^{*} - \epsilon$.

Proof: The results follow by setting $\mathcal{D} = \{d\}$ and $\mathcal{D} = \prod_{s \in \mathcal{S}} \mathcal{A}(s)$ respectively. ■
Theorem 4 and part (b) of Corollary 2 for the special case of finite state and action spaces appears in Satia (1968) with an additional assumption that the set of conditional measures $\mathcal{P}(s,a)$ is convex. (Their proof, in fact, extends to non-convex $\mathcal{P}(s,a)$.) Also, they do not explicitly prove that the solution of (30) is indeed the robust value function. Theorem 5 for general $\mathcal{D}$, and in particular for $\mathcal{D} = \{d\}$, is new. The special case $\mathcal{D} = \{d\}$ is crucial for establishing the policy improvement algorithm.

From Theorem 5, Corollary 2 and convergence results for contraction operators on Banach spaces, it follows that the robust value iteration algorithm displayed in Figure 1 computes an ϵ-optimal policy. This algorithm is the robust analog of the value iteration algorithm for non-robust DPs (see Section 6.3.2 in Puterman, 1994, for details). The following Lemma establishes this approximation result for the robust value iteration algorithm.

Lemma 1 *Let $\tilde{V}$ be the output of the robust value iteration algorithm shown in Figure 1. Then*

$$\|\tilde{V} - V^*_\lambda\| \leq \frac{\epsilon}{4},$$

where V^*_λ is the optimal value defined in (20). Let d be the decision rule

$$\inf_{p \in \mathcal{P}(s,d(s))} \mathbf{E}^p\big[r(s,d(s),s') + \lambda\tilde{V}(s')\big] \geq \sup_{a \in \mathcal{A}(s)} \Big\{ \inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p\big[r(s,a,s') + \lambda\tilde{V}(s')\big]\Big\} - \frac{\epsilon}{2}.$$

Then, the policy $\pi = (d, d, \ldots)$ is ϵ-optimal.

Proof: Since Lemma 5 establishes that $\mathcal{L}_D$ is a contraction operator, this result is a simple extension of Theorem 6.3.1 in Puterman (1994). The details are left to the reader. ■

Suppose the action set $\mathcal{A}(s)$ is finite. Then robust value iteration reduces to

$$\tilde{V}(s) = \max_{a \in \mathcal{A}(s)} \Big\{ \inf_{p \in \mathcal{P}_n(s,a)} \mathbf{E}^p\Big[r(s,a,s') + V_{n+1}(s')\Big]\Big\}.$$

For this iteration to be efficient one must be able to efficiently solve the optimization problem $\inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p[v]$ for a specified $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$ and $v \in \mathbf{R}^{|\mathcal{S}|}$. These optimization problems are identical to those solved in finite state problems. In Section 4 we show that for suitable choices for the set $\mathcal{P}(s,a)$ of conditional measures the complexity of solving such problems is only modestly larger than evaluating $\mathbf{E}^p[v]$ for a fixed p.

We next present a policy iteration approach for computing V^*_λ. As a first step, Lemma 2 below establishes that policy evaluation is a robust optimization problem.

Lemma 2 (Policy evaluation) *Let d be a deterministic decision rule and $\pi = (d, d, \ldots)$ be the corresponding stationary policy. Then V^π is the optimal solution of the robust optimization problem*

$$\begin{array}{ll} \textit{maximize} & \sum_{s \in \mathcal{S}} \alpha(s)V(s), \\ \textit{subject to} & V(s) \leq \mathbf{E}^p[r_s + \lambda V], \quad \forall p \in \mathcal{P}(s,d(s)), s \in \mathcal{S}, \end{array} \tag{31}$$

where $\alpha(s) > 0$, $s \in \mathcal{S}$, and $r_s \in \mathcal{R}^{|\mathcal{S}|}$ with $r_s(s') = r(s,d(s),s')$, $s' \in \mathcal{S}$.

Proof: The constraint in (31) can be restated as $V \leq \mathcal{L}_d V$, where $\mathcal{L}_d = \mathcal{L}_\mathcal{D}$ with $\mathcal{D} = \{d\}$. Corollary 2 implies that $V^\pi = \mathcal{L}_d V^\pi$, i.e. V^π is feasible for (31). Therefore, the optimal value of (31) is at least $\sum_{s \in \mathcal{S}} \alpha(s)V^\pi(s)$.

For every $s \in \mathcal{S}$, choose $p_s \in \mathcal{P}(s,d(s))$ such that

$$V^\pi(s) = \mathcal{L}_d V^\pi(s) \geq \mathbf{E}^{p_s}[r(s,d(s),s') + \lambda V^\pi(s')] - \epsilon.$$

Then for any V feasible for (31)

$$\begin{aligned} V(s) - V^\pi(s) &\leq \mathbf{E}^{p_s}[r(s,d(s),s') + \lambda V(s')] - \Big(\mathbf{E}^{p_s}[r(s,d(s),s') + \lambda V^\pi(s')] - \epsilon\Big), \\ &= \lambda \mathbf{E}^{p_s}\big[V(s') - V^\pi(s')\big] + \epsilon. \end{aligned}$$

Iterating this argument for n time steps, we get the bound

$$V(s) - V^\pi(s) \leq \lambda^n \|V - V^\pi\| + \frac{\epsilon}{1-\lambda}.$$

Since n and ϵ are arbitrary, all V feasible for (31) satisfy $V \leq V^\pi$. Since $\alpha(s) > 0$, $s \in \mathcal{S}$, it follows that the value of (31) is at most $\sum_{s \in \mathcal{S}} \alpha(s)V^\pi(s)$. This establishes the result. ■

```
The Robust Policy Iteration Algorithm:
Input: decision rule $d_0$, $\epsilon > 0$
Output: $\epsilon$-optimal decision rule $d^*$
Set $n = 0$ and $\pi_n = (d_n, d_n, \ldots)$. Solve (31) to compute $V^{\pi_n}$. Set $\tilde{V} \leftarrow \mathcal{L}_{\mathcal{D}} V^{\pi_n}$, $\mathcal{D} = \prod_{s\in\mathcal{S}} \mathcal{A}(s)$
For each $s \in \mathcal{S}$, choose
        $d_{n+1}(s) \in \Big\{a \in \mathcal{A}(s) : \inf_{p\in\mathcal{P}(s,a)} \mathbf{E}^p\big[r(s,a,s') + \lambda V(s')\big] \geq \tilde{V}(s) - \epsilon\Big\};$
    setting $d_{n+1}(s) = d_n(s)$ if possible.
while $\big(d_{n+1} \neq d_n\big)$ do
        $n = n + 1$; Solve (31) to computer $V^{\pi_n}$. Set $\tilde{V} \leftarrow \mathcal{L}_{\mathcal{D}} V^{\pi_n}$, $\mathcal{D} = \prod_{s\in\mathcal{S}} \mathcal{A}(s)$
        For each $s \in \mathcal{S}$, choose
            $d_{n+1}(s) \in \Big\{a \in \mathcal{A}(s) : \inf_{p\in\mathcal{P}(s,a)} \mathbf{E}^p\big[r(s,a,s') + \lambda V(s')\big] \geq \tilde{V}(s) - \epsilon\Big\};$
        setting $d_{n+1}(s) = d_n(s)$ if possible.
end while
return $d_{n+1}$
```

Figure 2: Robust policy iteration algorithm

Since $\mathbf{E}^p[r_s + \lambda V]$ is a linear function of p, (31) is a convex optimization problem. Typically, (31) can be solved efficiently only if $\mathcal{S}$ is finite and the robust constraint can be reformulated as a small collection of deterministic constraints. In Section 4 we introduce some natural candidates for the set $\mathcal{P}(s, a)$ of conditional measures. Dualizing the constraints in (31) leads to a compact representation for some of these sets. However, for most practical applications, the policy evaluation step is computationally expensive and is usually replaced by a m-step look-ahead value iteration (Puterman, 1994).

Lemma 1 leads to the robust policy iteration algorithm displayed in Figure 2. Suppose (31) is efficiently solvable; then finite convergence of this algorithm for the special case of finite state and action spaces follows from Theorem 6.4.2 in Puterman (1994). A rudimentary version of robust policy iteration algorithm for this special case appears in Satia and Lave (1973) (see also Satia, 1968). They compute the value of a policy $\pi = (d, d, \ldots)$, i.e. solve the robust optimization problem (31), via the following iterative procedure:

(a) For every $s \in \mathcal{S}$, fix $p_s \in \mathcal{P}(s, d(s))$. Solve the set of equations

$$V(s) = \mathbf{E}^{p_s}[r(s, d(s), s') + \lambda V(s')], \quad s \in \mathcal{S}.$$

Since $\lambda < 1$, these set of equations has a unique solution (see Theorem 6.1.1 in Puterman, 1994).

(b) Fix V, and solve

$$\tilde{p}(s) \leftarrow \operatorname*{argmin}_{p\in\mathcal{P}(s,d(s))} \Big\{\mathbf{E}^p[r(s, d(s), s') + \lambda V(s')]\Big\}, \quad s \in \mathcal{S}.$$

If $V(s) = \mathbf{E}^{\tilde{p}_s}[r(s, d(s), s') + \lambda V(s')]$, for all $s \in \mathcal{S}$, stop; otherwise, $p(s) \leftarrow \tilde{p}(s)$, $s \in \mathcal{S}$, return to (a).

However, it is not clear, and Satia and Lave (1973) do not show, that this iterative procedure converges.

Given the relative ease with which value iteration and policy iteration translate to the robust setting, one might attempt to solve the robust DP by the following natural analog of the linear programming method for DP (Puterman, 1994):

$$\begin{array}{rl} \text{maximize} & \sum_{s\in\mathcal{S}} \alpha(s)V(s), \\ \text{subject to} & V(s) \geq \inf_{p\in\mathcal{P}(s,a)} \mathbf{E}^{p}[r(s,a,s') + \lambda V(s')], \quad a \in \mathcal{A}(s), s \in \mathcal{S}. \end{array} \tag{32}$$

Unfortunately, (32) is not a convex optimization problem. Hence, the LP method does not appear to have a tractable analog in the robust setting.

Recall that in the beginning of this section we had proposed two models for the adversary. The first was a *dynamic* model where the measures $\mathcal{T}^{\pi}$ consistent with a policy π satisfies Rectangularity. So far we have assumed that this model prevails. In the second, *static* model, the adversary was restricted to employing a fixed $p_{sa} \in \mathcal{P}(s,a)$ whenever the state-action pair (s,a) is encountered. The last result in this section establishes that if the decision maker is restricted to stationary policies the implications of the static and dynamic models are, in fact, identical.

Lemma 3 (Dynamic vs Static adversary) *Let d be any decision rule and let $\pi = (d, d, \ldots)$ be the corresponding stationary policy. Let V_{λ}^{π} and $\widehat{V}_{\lambda}^{\pi}$ be the value of the π in the dynamic and static model respectively. Then $\widehat{V}_{\lambda}^{\pi} = V_{\lambda}^{\pi}$.*

Proof: We prove the result for deterministic decision rules. The same technique extends to randomized policies but the notation becomes complicated.

Clearly $\widehat{V}_{\lambda}^{\pi} \geq V_{\lambda}^{\pi}$. Thus, we only need to establish that $\widehat{V}_{\lambda}^{\pi} \leq V_{\lambda}^{\pi}$. Fix $\epsilon > 0$ and choose $\bar{p} : \mathcal{S} \mapsto \mathcal{M}(\mathcal{S})$ such that $\bar{\mathbf{p}}_s \in \mathcal{P}(s, d(s))$, for all $s \in \mathcal{S}$, and $V_{\lambda}^{\pi}(s) \geq \mathbf{E}^{\bar{\mathbf{p}}_s}[r(s,d(s),s') + \lambda V_{\lambda}^{\pi}(s')] - \epsilon$. Let $V_{\lambda\bar{\mathbf{p}}}^{\pi}$ denote the non-robust value of the policy π corresponding to the fixed conditional measure $\bar{\mathbf{p}}$. Clearly $V_{\lambda\bar{\mathbf{p}}}^{\pi} \geq \widehat{V}_{\lambda}^{\pi}$. Thus, the result will follow if we show that $V_{\lambda\bar{\mathbf{p}}}^{\pi} \leq V_{\lambda}^{\pi}$.

From results for non-robust DP we have that $V_{\lambda\bar{\mathbf{p}}}^{\pi} = \mathbf{E}^{\bar{\mathbf{p}}_s}[r(s,d(s),s') + \lambda V_{\lambda\bar{\mathbf{p}}}^{\pi}(s')]$. Therefore,

$$\begin{array}{rcl} V_{\lambda\bar{\mathbf{p}}}^{\pi} - V_{\lambda}^{\pi}(s) & \leq & \Big(\mathbf{E}^{p_s}[r(s,a,s') + \lambda V_{\lambda\bar{\mathbf{p}}}^{\pi}(s')]\Big) - \Big(\mathbf{E}^{p_s}[r(s,d(s),s') + \lambda V_{\lambda}^{\pi}(s')] - \epsilon\Big), \\ & = & \lambda \mathbf{E}^{p_s}\big[\widehat{V}_{\lambda}^{\pi}(s') - V_{\lambda}^{\pi}(s')\big] + \epsilon. \end{array}$$

Iterating this bound for n time steps, we get

$$V_{\lambda\bar{\mathbf{p}}}^{\pi}(s) - V_{\lambda}^{\pi}(s) \leq \lambda^n \|\widehat{V}_{\lambda\bar{\mathbf{p}}}^{\pi} - V_{\lambda}^{\pi}\| + \frac{\epsilon}{1-\lambda}.$$

Since n and ϵ are arbitrary, it follows that $V_{\lambda\bar{\mathbf{p}}}^{\pi} \leq V_{\lambda}^{\pi}$. ■

In the proof of the result we have implicitly established that the "best-response" of dynamic adversary when the decision maker employs a stationary policy is, in fact, static, i.e. the adversary chooses the same $p_{sa} \in \mathcal{P}(s,a)$ every time the pair (s,a) is encountered. Consequently, the optimal stationary policy in a static model can be computed by solving (30). Bagnell et al. (2001) establish that when the set $\mathcal{P}(s,a)$ of conditional measures is convex and the decision maker is restricted to stationary policies the optimal policies for the decision maker and the adversary is the same in both the static and dynamic models. We extend this result to non-convex sets. In addition we show that the value of any stationary policy, optimal or otherwise, is the same in both models. While solving (30) is, in general, NP-complete (Littman, 1994), the problem is tractable provided the sets are $\mathcal{P}(s,a)$ are "nice" convex sets. In particular, the problem is tractable for the families of sets discussed in Section 4.

Lemma 3 highlights an interesting asymmetry between the decision maker and the adversary that is a consequence of the fact that the adversary plays second. While it is optimal for a dynamic adversary to play static (stationary) policies when the decision maker is restricted to stationary policies, it is not optimal for the decision maker to play stationary policies against a static adversary. The optimal policy for the decision maker in the static model are the so-called universal policy (Cover, 1991).

4 Tractable sets of conditional measures

Section 2 and Section 3 were devoted to extending results from non-robust DP theory. In this and the next section we focus on computational issues. Since computations are only possible when state and action spaces are finite (or are suitably truncated versions of infinite sets), we restrict ourselves to this special case. The results in this section are not new and are included for completeness. They were first obtained by El Ghaoui and Nilim Nilim and El Ghaoui (2002).

In the absence of any ambiguity, the value of an action $a \in \mathcal{A}(s)$ in state $s \in \mathcal{S}$ is given by $\mathbf{E}^p[v] = p^T v$, where p is the conditional measure and v is a random variable that takes value $v(s') = r(s, a, s') + V(s')$ in state $s' \in \mathcal{S}$. Thus, the complexity of evaluating the value of a state-action pair is $\mathcal{O}(|\mathcal{S}|)$. When the conditional measure is ambiguous, the value of the state-action pair is (s, a) is given by $\inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p[v]$. In this section, we introduce three families of sets of conditional measures $\mathcal{P}(s, a)$ which only result in a modest increase in complexity, typically logarithmic in $|\mathcal{S}|$. These families of sets are constructed from approximations of the confidence regions associated with density estimation. Two of these families are also discussed in Nilim and El Ghaoui (2002). We distinguish our contribution in the relevant sections.

Note that since $\sup_{p \in \mathcal{P}(s,a)} \mathbf{E}^p[v] = -\inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p[-v]$, it follows that the recursion (18) for the optimistic value function can also be computed efficiently for these families of sets.

4.1 Sets based on relative entropy

As mentioned in the introduction, the motivation for the robust methodology was to systematically correct for the statistical errors associated with estimating the transition probabilities using historical data. Thus, a natural choice for the sets $\mathcal{P}(s, a)$ of conditional measures are the confidence regions associated with density estimation. In this section, we show how to construct such sets for any desired confidence level $\omega \in (0, 1)$. We also show that the optimization problem $\inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p[v]$ can be efficiently solved for this class of sets.

Suppose the underlying controlled Markov chain is stationary. Suppose also that we have historical data consisting of triples $\{(s_j, a_j, s'_j) : j \geq 1\}$, with the interpretation that state s'_j was observed in period $t + 1$ when the action a_j was employed in state s_j in period t. Then the maximum likelihood estimate $\widehat{p}_{sa}$ of the conditional measure corresponding to the state-action pair (s, a) is given by

$$\widehat{p}_{sa} = \operatorname*{argmax}_{p \in \mathcal{M}(\mathcal{S})} \Big\{ \sum_{s' \in \mathcal{S}} n(s'|s, a) \log(p(s')) \Big\}, \tag{33}$$

where

$$n(s'|a, s) = \sum_j \mathbf{1}\Big((s, a, s') = (s_j, a_j, s'_j)\Big),$$

is the number of samples of the triple (s, a, s'). Let $q \in \mathcal{M}(\mathcal{S})$ be defined as

$$q(s') = \frac{n(s'|s, a)}{\sum_{u \in \mathcal{S}} n(u|s, a)}, \quad s' \in \mathcal{S}.$$

Then, (33) is equivalent to

$$\widehat{p}_{sa} = \operatorname*{argmin}_{p \in \mathcal{M}(\mathcal{S})} D(q\|p), \tag{34}$$

where $D(p_1\|p_2)$ is the Kullback-Leibler or the relative entropy distance (see Chapter 2 in Cover and Thomas, 1991) between two measures $p_1, p_2 \in \mathcal{M}(\mathcal{S})$ and is defined as follows:

$$D(p_1\|p_2) = \sum_{s\in\mathcal{S}} p_1(s)\log\Big(\frac{p_1(s)}{p_2(s)}\Big). \tag{35}$$

The function $D(p_1\|p_2) \geq 0$ with equality if and only if $p_1 = p_2$ (however, $D(p_1\|p_2) \neq D(p_2\|p_1)$). Thus, we have that the maximum likelihood estimate of the conditional measure is given by

$$\widehat{p}_{sa}(s') = q(s') = \frac{n(s'|s,a)}{\sum_{u\in\mathcal{S}} n(u|s,a)}, \quad s', s \in \mathcal{S},\ a \in \mathcal{A}(s). \tag{36}$$

More generally, let $g^j : \mathcal{S} \mapsto \mathbf{R}$, $j = 1, \ldots, k$ be k functions defined on the state space $\mathcal{S}$ (typically, $g^j(s) = s^j$, i.e. j-th moment) and

$$\bar{g}^j_{sa} = \frac{1}{n_{sa}} \sum_{s\in\mathcal{S}} n(s'|s,a) g^j(s'), \quad j = 1, \ldots, k,$$

be the sample averages of the moments corresponding to the state-action pair (s,a). Let $p^0_{sa} \in \mathcal{M}(\mathcal{S})$ be the prior distribution on $\mathcal{S}$ conditioned on the state-action pair (s,a). Then the maximum likelihood solution $\widehat{p}_{sa}$ is given by

$$\widehat{p}_{sa} = \underset{\{p\in\mathcal{M}(\mathcal{S}):\mathbf{E}^p[g^j]=\bar{g}^j_{sa}, j=1,\ldots,k\}}{\operatorname{argmin}} D(p\|p^0) \tag{37}$$

provided the set $\big\{p \in \mathcal{M}(\mathcal{S}) : \mathbf{E}^p[g^j] = \bar{g}^j_{sa}, j = 1, \ldots, k\big\} \neq \emptyset$.

Let p_{sa}, $a \in \mathcal{A}(s), s \in \mathcal{S}$ denote the unknown *true* state transition of the stationary Markov chain. Then a standard result in statistical information theory (see Cover and Thomas, 1991, for details) implies the following convergence in probability:

$$n_{sa} D(p_{sa}\|\widehat{p}_{sa}) \Longrightarrow \frac{1}{2}\chi^2_{|S|-1}, \tag{38}$$

where $n_{sa} = \sum_{s'\in\mathcal{S}} n(s'|s,a)$ is the number of samples of the state-action pair (s,a) and $\chi^2_{|S|-1}$ denotes a χ^2 random variable with $|S|-1$ degrees of freedom (note that the maximum likelihood estimate $\widehat{p}_{sa}$ is, itself, a function of the sample size n_{sa}). Therefore,

$$\begin{aligned}\mathbf{P}\Big\{p : D(p\|\widehat{p}_{sa}) \leq t\Big\} &\approx \mathbf{P}\Big\{\chi^2_{|S|-1} \leq 2n_{sa}t\Big\},\\ &= \mathcal{F}_{|\mathcal{S}|-1}(2n_{sa}t).\end{aligned}$$

Let $\omega \in (0,1)$ and $t_\omega = \mathcal{F}^{-1}_{|\mathcal{S}|-1}(\omega)/(2n_{sa})$. Then

$$\mathcal{P} = \Big\{p \in \mathcal{M}(\mathcal{S}) : D(p\|\widehat{p}_{sa}) \leq t_\omega\Big\}, \tag{39}$$

is the ω-confidence set for the true transition probability p_{sa}. Since $D(p\|q)$ is a convex function of the pair (p,q) (Cover and Thomas, 1991), $\mathcal{P}$ is convex for all $t \geq 0$.

The following results establish that an ϵ-approximate solution for the robust problem corresponding to the set $\mathcal{P}$ in (39) can be computed efficiently.

Lemma 4 *The value of optimization problem :*

$$\begin{aligned}&\textit{minimize} \quad \mathbf{E}^p[v]\\ &\textit{subject to} \quad p \in \mathcal{P} = \Big\{p \in \mathcal{M}(\mathcal{S}) : D(p\|q) \leq t, q \in \mathcal{M}(\mathcal{S})\Big\},\end{aligned} \tag{40}$$

where $t > 0$, is equal to

$$-\min_{\gamma\geq 0}\Big\{t\gamma+\gamma\log\Big(\mathbf{E}^{q}\big[\exp\big(-\frac{v}{\gamma}\big)\big]\Big)\Big\}. \tag{41}$$

The complexity of computing an ϵ-optimal solution for (41) is $\mathcal{O}\Big(|S|\,\Big\lceil\log_2\frac{\Delta v\max\{t,|t+\log(q_{\min})|\}}{2\epsilon t}\Big\rceil\Big)$, where $\Delta v = \max_{s\in\mathcal{S}}\{v(s)\} - \min_{s\in\mathcal{S}}\{v(s)\}$ and $q_{\min} = \mathbf{P}(v(s) = \min\{v\})$.

Proof: The Lagrangian $\mathcal{L}$ for the optimization problem (40) is given by

$$\mathcal{L} = \sum_{s\in\mathcal{S}} p(s)v(s) - \gamma\Big(t - \sum_{s\in\mathcal{S}} p(s)\log\Big(\frac{p(s)}{q(s)}\Big)\Big) - \mu\Big(\sum_{s\in\mathcal{S}} p(s) - 1\Big), \tag{42}$$

where $\gamma \geq 0$ and $\mu \in \mathbf{R}$. Taking the derivative of $\mathcal{L}$ with respect to $p(s)$ and setting it to zero, we get

$$v(s) + \gamma\Big(\log\Big(\frac{p(s)}{q(s)}\Big) + 1\Big) - \mu = 0, \qquad s \in \mathcal{S},$$

i.e.

$$\gamma\log\Big(\frac{p(s)}{q(s)}\Big) + v(s) = \mu - \gamma. \tag{43}$$

From (43) it follows that

$$p(s) = q(s)\exp\Big(-1+\frac{\mu - v(s)}{\gamma}\Big), \quad s \in \mathcal{S}. \tag{44}$$

Thus, the non-negative constraints $p(s) \geq 0$ are never active. Since p is constrained to be a probability, i.e. $\sum_{s\in\mathcal{S}} p(s) = 1$, (43) implies that the Lagrangian

$$\mathcal{L} = \mu - \gamma - \gamma t.$$

Also, (44) together with the fact that $p \in \mathcal{M}(\mathcal{S})$ implies that

$$\mu - \gamma = -\gamma\log\Big(\sum_{s\in\mathcal{S}} q(s)e^{\frac{v(s)}{\gamma}}\Big)$$

Thus, the Lagrangian $\mathcal{L}(\gamma) = -t\gamma - \gamma\log\Big(\mathbf{E}^{q}\big[\exp\big(\frac{v}{\gamma}\big)\big]\Big)$. For $t > 0$ the set $\mathcal{P}$ in (40) has a strictly feasible point; therefore, the value of (40) is equal to $\max_{\gamma\geq 0}\mathcal{L}(\gamma)$.

Suppose $v(s) = v$ for all $s \in \mathcal{S}$. Then, the value of (40) is trivially v. Next assume that there exist $s, s' \in \mathcal{S}$ such that $v(s) \neq v(s')$. In this case, by suitably shifting and scaling the vector v, one can assume that $v(s) \in [0, 1]$, $\min_{s\in\mathcal{S}}\{v(s)\} = 0$ and $\max_{s\in\mathcal{S}}\{v(s)\} = 1$. Note that this shifting and scaling is an $\mathcal{O}(|\mathcal{S}|)$ operation.

Let $f(\gamma) = \gamma t + \gamma\log\big(\sum_{s\in\mathcal{S}} q(s)e^{-\frac{v(s)}{\gamma}}\big)$ denote the objective function of (41). Then $f(\gamma)$ is convex and

$$f'(\gamma) = t + \log\Big(\sum_{s\in\mathcal{S}} q(s)e^{-\frac{v(s)}{\gamma}}\Big) + \frac{1}{\gamma}\frac{\sum_{s\in\mathcal{S}} q(s)v(s)e^{-\frac{v(s)}{\gamma}}}{\sum_{s\in\mathcal{S}} q(s)e^{-\frac{v(s)}{\gamma}}}.$$

Since $f(\gamma)$ is convex, it follows that $f'(\gamma)$ is non-decreasing. It is easy to verify that $f'(0) = \lim_{\gamma\to 0} f'(\gamma) = t + \log(q_{\min})$, where $q_{\min} = \text{Prob}(v = 0)$, $f'(\frac{1}{t}) > 0$ and $|f'(\gamma)| \leq \max\big\{t, |t + \log(q_{\min})|\big\}$.

Clearly, $\gamma = 0$ is optimum if $f'(0) \geq 0$. Otherwise, the optimum value lies in the interval $[0, \frac{1}{t}]$, and after N iterations of a bisection algorithm the optimum value γ^* is guaranteed to lie in an interval $[\gamma_1, \gamma_2]$ with $\gamma_2 - \gamma_1 \leq \frac{1}{t}2^{-N}$. Let $\bar{\gamma} = \frac{1}{2}(\gamma_1 + \gamma_2)$. Then

$$\begin{aligned} f(\gamma^*) - f(\bar{\gamma}) &\geq -\frac{\epsilon}{2}\left|f'(\bar{\gamma})\right|, \\ &\geq -\frac{\epsilon}{2}\max\big\{t, |t+\log(q_{\min})|\big\}. \end{aligned} \tag{45}$$

Thus, it follows that an ϵ-optimal solution of $\min_\gamma f(\gamma)$ can be computed in $\lceil \log_2 \frac{\max\{t, |t+\log(q_{\min})|\}}{2\epsilon t} \rceil$ bisections. The result follows by recognizing that each evaluation of $f'(\gamma)$ is an $\mathcal{O}(|\mathcal{S}|)$ operation. ■

As mentioned above, relative entropy-based sets of conditional measures of the form (39) were first introduced in Nilim and El Ghaoui (2002). Our analysis is different but essentially equivalent to their approach.

4.2 Sets based on $\mathcal{L}_2$ approximations for the relative entropy

In this section we consider conservative approximations for the relative entropy sets. For this family of sets the optimization $\inf_{p\in\mathcal{P}} \mathbf{E}^p[v]$ can solved to optimality in $\mathcal{O}(|\mathcal{S}|\log(|\mathcal{S}|))$ time.

Since $\log(1+x) \le x$ for all $x \in \mathbf{R}$, it follows that

$$D(p\|q) = \sum_{s\in\mathcal{S}} p(s)\log\Big(\frac{p(s)}{q(s)}\Big) \le \sum_{s\in c\mathcal{S}} \Big(p(s)\cdot\frac{p(s)-q(s)}{q(s)}\Big) = \sum_{s\in\mathcal{S}} \frac{(p(s)-q(s))^2}{q(s)}.$$

Thus, a conservative approximation for the uncertainty set defined in (39) is given by

$$\mathcal{P} = \Big\{p \in \mathcal{M}(\mathcal{S}) : \sum_{s\in\mathcal{S}} \frac{(p(s)-q(s))^2}{q(s)} \le t\Big\}. \tag{46}$$

Lemma 5 *The value of optimization problem :*

$$\begin{array}{ll} \textit{minimize} & \mathbf{E}^p[v] \\ \textit{subject to} & p \in \mathcal{P} = \Big\{p \in \mathcal{M}(\mathcal{S}) : \sum_{s\in\mathcal{S}} \frac{(p(s)-q(s))^2}{q(s)} \le t\Big\}, \end{array} \tag{47}$$

is equal to

$$\max_{\mu\ge\mathbf{0}} \Big\{\mathbf{E}^q[v-\mu] - \sqrt{t\mathbf{Var}^q[v-\mu]}\Big\}, \tag{48}$$

and the complexity of (48) is $\mathcal{O}(|\mathcal{S}|\log(|\mathcal{S}|))$.

Proof: Let $y = p - q$. Then $p \in \mathcal{P}$ if and only if $\sum_s \frac{y^2(s)}{q(s)} \le t$, $\sum_s y(s) = 0$ and $y \ge -q$. Thus, the value of (47) is equal to

$$\begin{array}{rll} \mathbf{E}^q[v] + & \text{minimize} & \sum_s y(s)v(s), \\ & \text{subject to} & \sum_s \frac{y^2(s)}{q(s)} \le t, \\ & & \sum_s y(s) = 0, \\ & & y \ge -q. \end{array} \tag{49}$$

Lagrangian duality implies that the value of (49) is equal to

$$\begin{aligned} & \mathbf{E}^q[v] + \max_{\mu\ge\mathbf{0},\gamma\ge0} \min_{\left\{y:\sum_s \frac{y^2(s)}{q(s)}\le t\right\}} \Big\{-\sum_{s\in\mathcal{S}}\mu(s)q(s) + \sum_{s\in\mathcal{S}} y(s)\big(v(s)-\gamma-\mu(s)\big)\Big\} \\ = \; & \max_{\mu\ge\mathbf{0},\gamma\ge0} \Big\{\mathbf{E}^q[v-\mu] - \sqrt{t\sum_{s\in\mathcal{S}} q(s)(v(s)-\mu(s)-\gamma)^2}\Big\}, \qquad (50) \\ = \; & \max_{\mu\ge\mathbf{0}} \Big\{\mathbf{E}^q[v-\mu] - \sqrt{t\mathbf{Var}^q[v-\mu]}\Big\}. \qquad (51) \end{aligned}$$

This establishes that the value of (47) is equal to that of (48).

The optimum value of the inner minimization in (50) is attained at

$$y^*(s) = -\frac{\sqrt{tq(s)}z(s)}{\|\mathbf{z}\|}, \quad s \in \mathcal{S},$$

where $z(s) = \sqrt{q(s)}\big(v(s) - \mu(s) - \mathbf{E}^q[v - \mu]\big)$, $s \in \mathcal{S}$. Let $\mathcal{B} = \big\{s \in \mathcal{S} : \mu(s) > 0\big\}$. Then complementary slackness conditions imply that $y^*(s) = -q(s)$, for all $s \in \mathcal{B}$, or equivalently,

$$v(s) - \mu(s) = \frac{\|\mathbf{z}\|}{\sqrt{t}} + \mathbf{E}^q[v - \mu] = \alpha, \quad \forall s \in \mathcal{B}, \tag{52}$$

i.e. $v(s) - \mu(s)$ is a constant for all $s \in \mathcal{B}$. Since the optimal value of (47) is at least $v_{\min} = \min_{s\in\mathcal{S}}\{v(s)\}$, it follows that $\alpha \geq v_{min}$.

Suppose α is known. Then the optimal μ^* is given by

$$\mu^*(s) = \begin{cases} v(s) - \alpha, & v(s) \geq \alpha, \\ 0, & \text{otherwise.} \end{cases} \tag{53}$$

Thus, dual optimization problem (48) reduces to solving for the optimal α. To this end, let $\{\widehat{v}(k) : 1 \leq k \leq |\mathcal{S}|\}$ denote the values $\{v(s) : s \in \mathcal{S}\}$ arranged in increasing order – an $\mathcal{O}(|\mathcal{S}|\log(|\mathcal{S}|))$ operation. Let $\widehat{q}$ denote the sorted values of the measure q.

Suppose $\alpha \in [\widehat{v}_n, \widehat{v}_{n+1})$. Then

$$\mathbf{E}^q[v - \mu] = a_n + b_n\alpha, \quad \mathbf{Var}^q[v - \mu] = c_n + b_n\alpha^2 + (a_n + b_n\alpha)^2,$$

where $a_n = \sum_{k\leq n}\widehat{q}(k)\widehat{v}(k)$, $b_n = \sum_{k>n}\widehat{q}(k)$ and $c_n = \sum_{k\leq n}\widehat{q}(k)\widehat{v}^2(k)$. Note that, once the sorting is done, computing $\{(a_n, b_n, c_n) : 1 \leq n \leq |\mathcal{S}|\}$ is $\mathcal{O}(|\mathcal{S}|)$.

The dual objective $f(\alpha)$ as a function of α is

$$\begin{aligned} f(\alpha) &= \mathbf{E}^q[v - \mu] - \sqrt{t\mathbf{Var}^q[v - \mu]}, \\ &= a_n + b_n\alpha - \sqrt{t(c_n + b_n\alpha^2 - (a_n + b_n\alpha)^2)}. \end{aligned}$$

If α is optimal, it must be that $f'(\alpha) = 0$, i.e. α is root of the quadratic equation

$$b_n^2\big(c_n + b_n\alpha^2 - (a_n + b_n\alpha)^2\big) = t\big(b_n(1 - b_n)\alpha - a_n\big)^2 \tag{54}$$

Thus, the optimal α can be computed by sequentially checking whether a root of (54) lies in $[v_n, v_{n+1})$, $n = 1, \ldots, |\mathcal{S}|$. Since this is an $\mathcal{O}(|\mathcal{S}|)$ operation, we have that the overall complexity of computing a solution of (48) is $\mathcal{O}(|\mathcal{S}|\log(|\mathcal{S}|))$. ■

Sets of conditional measures of the form (46) have also been investigated in Nilim and El Ghaoui (2002). However, they do not identify these sets as inner, i.e. conservative, approximations of relative entropy sets. Moreover, they are not able to solve the problem (47) – their algorithm is only able to solve (47) when the set (46) is expanded to $\{p : \mathbf{1}^T p = 1, \sum_{s\in\mathcal{S}}\frac{(p(s)-q(s))^2}{q(s)} \leq t\}$, i.e when the constraint $p \geq 0$ is dropped.

4.3 Sets based on $\mathcal{L}_1$ approximation to relative entropy

From Lemma 12.6.1 in Cover and Thomas (1991), we have that

$$D(p\|q) \geq \frac{1}{2\ln(2)}\|p - q\|_1^2,$$

where $\|p - q\|_1$ is the $\mathcal{L}_1$-distance between the measures $p, q \in \mathcal{M}(\mathcal{S})$. Thus, the set

$$\mathcal{P} = \Big\{p : \|p - q\|_1 \leq \sqrt{2\ln(2)t}\Big\}, \tag{55}$$

is an outer approximation, i.e. relaxation, of the relative entropy uncertainty set (39). To the best of our knowledge, this family of sets has not been previously analyzed.

Lemma 6 *The value of optimization problem :*

$$\begin{array}{ll} \textit{minimize} & \mathbf{E}^p[v] \\ \textit{subject to} & p \in \mathcal{P} = \Big\{ p : \|p - q\|_1 \leq \sqrt{2\ln(2)t} \Big\}, \end{array} \tag{56}$$

is equal to

$$\mathbf{E}^q[v] - \frac{1}{2}\Big(\sqrt{2\ln(2)t}\Big) \min_{\mu \geq \mathbf{0}} \Big\{ \big(\max_s \{v(s) - \mu(s)\} - \min_s \{v(s) - \mu(s)\} \big) \Big\}, \tag{57}$$

and the complexity of (57) is $\mathcal{O}(|\mathcal{S}| \log(|\mathcal{S}|))$.

Proof: Let $y(s) = (p(s) - q(s))$, $s \in \mathcal{S}$. Then $p \in \mathcal{P}$ if and only if $\|y\|_1 \leq \sqrt{2\ln(2)t}$, $\sum_{s\in\mathcal{S}} y(s) = 0$ and $y \geq -q$. Therefore, the value of (56) is equal to

$$\begin{array}{lll} \mathbf{E}^q[v] + & \text{minimize} & \sum_{s\in\mathcal{S}} y(s)v(s) \\ & \text{subject to} & \|y\|_1 \leq \sqrt{2\ln(2)t}, \\ & & \sum_{s\in\mathcal{S}} y(s) = 0, \\ & & y \geq -q. \end{array}$$

From Lagrangian duality we have that the value of this optimization problem is equal to

$$\begin{aligned} & \mathbf{E}^q[v] + \max_{\mu \geq \mathbf{0}, \gamma \in \mathbf{R}} \min_{y : \|y\|_1 \leq \sqrt{2\ln(2)t}} \Big\{ -\sum_{s\in\mathcal{S}} \mu(s)q(s) + \sum_{s\in\mathcal{S}} y(s)(v(s) - \mu(s) - \gamma) \Big\} \\ = \ & \mathbf{E}^q[v] + \max_{\mu \geq \mathbf{0}, \gamma \in \mathbf{R}} \Big\{ -\sum_{s\in\mathcal{S}} \mu(s)q(s) - \sqrt{2\ln(2)t}\|v - \mu - \gamma\mathbf{1}\|_\infty \Big\}, \\ = \ & \max_{\mu \geq \mathbf{0}} \Big\{ \mathbf{E}^q[v - \mu] - \frac{1}{2}\sqrt{2\ln(2)t}\big(\max_s \{v(s) - \mu(s)\} - \min_s \{v(s) - \mu(s)\} \big) \Big\}. \end{aligned}$$

Let μ^* be the optimal dual solution and let $\alpha = \max_{s\in\mathcal{S}}\{v(s) - \mu^*(s)\}$. It is easy to see that

$$\mu^*(s) = \begin{cases} v(s) - \alpha, & v(s) > \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, dual optimization problem (48) reduces to solving for the optimal α. To this end, let $\{\widehat{v}(k) : 1 \leq k \leq |\mathcal{S}|\}$ denote the values $\{v(s) : s \in \mathcal{S}\}$ arranged in increasing order – an $\mathcal{O}(|\mathcal{S}| \log(|\mathcal{S}|))$ operation. Let $\widehat{q}$ denote the sorted values of the measure q.

Suppose $\alpha \in [\widehat{v_n}, \widehat{v}_{n+1})$. Then, the dual function $f(\alpha)$ is given by

$$\begin{aligned} f(\alpha) &= \mathbf{E}^q[v - \mu] - \frac{1}{2}\sqrt{2\ln(2)t}\big(\max_s \{v(s) - \mu(s)\} - \min_s \{v(s) - \mu(s)\} \big), \\ &= \sum_{k \leq n} \widehat{q}(k)\widehat{v}(k) + \frac{1}{2}\sqrt{2t\ln(2)}\widehat{v}_1 + \Big(\sum_{k > n} \widehat{q}(k) - \sqrt{2\ln(2)t} \Big)\alpha. \end{aligned}$$

Since $f(\alpha)$ is linear, the optimal is always obtained at the end points. Thus, the optimal value of α is given by

$$\alpha = \min\Big\{ \widehat{v}(n) : \sum_{k > n} \widehat{q}(k) < \sqrt{2\ln(2)t} \Big\},$$

i.e. α can be computed in $\mathcal{O}(|\mathcal{S}|)$ time. ■

See Nilim and El Ghaoui (2002) for other families of sets, in particular sets based on $\mathcal{L}_\infty$ and $\mathcal{L}_2$ norms. Although these families are popular in modeling, they do not have any basis in statistical theory. Consequently, parameterizing these sets are nearly impossible. We, therefore, do not recommend using these sets to model ambiguity in the transition probability.

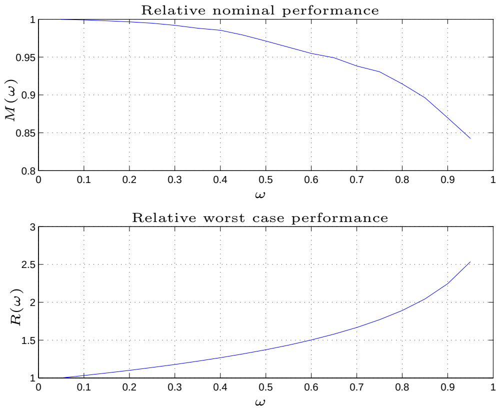

Figure 3: Robust stopping problem ($|\mathcal{S}| = 100$, $N = 10$, $m = 40$)

5 Computational results

5.1 Robust finite horizon optimal stopping problems

Suppose the state transition matrix $\mathbf{A}$ of a Markov chain where known. Clearly, the non-robust policy designed for $\mathbf{A}$ will then be superior to a robust policy designed for any set $\mathcal{P}$ containing $\mathbf{A}$. The rationale behind the robust formulation was that if there was an error in estimating $\mathbf{A}$ then the performance of the policy designed for $\mathbf{A}$ will be significantly worse than a robust policy. In this section we investigate this claim in the context of finite horizon optimal stopping problems.

In an optimal stopping problem the state evolves according to uncontrollable stationary Markov chain $\mathbf{M}$ on a finite state space $\mathcal{S}$. At each decision epoch t and each state $s \in \mathcal{S}$, the decision maker has two actions available: stop or continue. If the decision maker stops in state s at time t, the reward received is $g_t(s)$, and if the action is to continue, the cost incurred is $f_t(s)$; and the state s_{t+1} evolves according to $\mathbf{M}$. The problem has a finite time horizon N and, if the decision maker does not stop before N, the reward at time N is $h(s)$. Once stopped, the state remains in the stopped state, yielding no reward thereafter. The objective is to choose a policy to maximize the total expected reward.

The experimental setup was as follows. Once the time horizon N and the size of the state space $\mathcal{S}$ was selected, a transition matrix $\mathbf{A}$ was randomly generated. In order to keep the problem tractable we assumed a bound on the number m of 1-step neighbors and ensured that $\mathbf{A}$ induced an irreducible Markov chain. The rewards $g_t(s)$, the cost $f_t(s)$ and the terminal reward $h(s)$ were all randomly generated.

A single sample path of length $100\,|\mathcal{S}|^2$ was generated according to the above (randomly selected) Markov chain. This sample path was used to compute the maximum likelihood estimate $\mathbf{A}_{ml}$ of the transition matrix. We will call $\mathbf{A}_{ml}$ the nominal Markov chain. The non-robust DP assumed that the underlying Markov chain is governed by $\mathbf{A}_{ml}$. Let V_0^{nr} denote the non-robust value function. The ambiguity in the transition matrix was modeled by the relative entropy sets defined in (39). The ambiguity structure was applied independently to each row of the transition matrix. For each $\omega \in \{0.05, 0.01, \ldots, 0.95\}$ we computed the robust stopping policy using the robust Bellman recursion (11). Let $V_0^{r,\omega}$ denote the robust value function corresponding to the confidence level ω.

The first performance measure we consider is the loss associated with employing robust policies on the nominal Markov chain. Clearly the non-robust value function V_0^{nr} is optimal value of the optimal stopping problem defined on the nominal chain. Let $V_{0,ml}^{r,\omega}$ denote the reward generated by robust policy corresponding

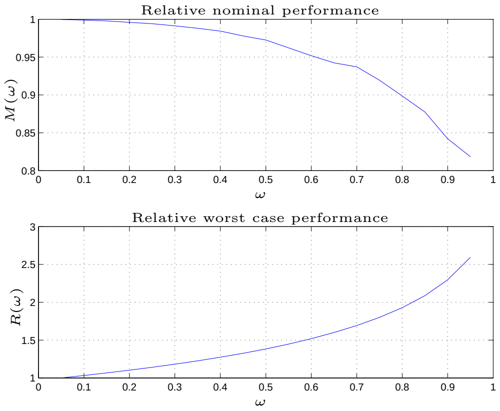

Figure 4: Robust stopping problem ($|\mathcal{S}| = 100$, $N = 10$, $m = 80$)

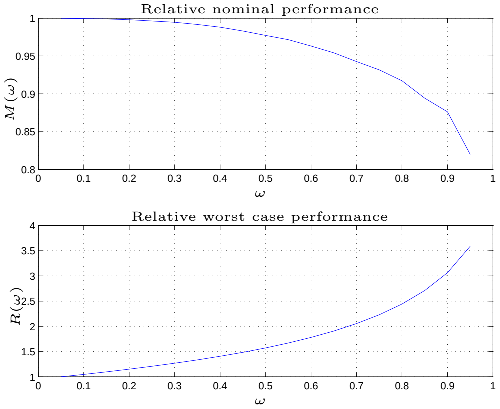

Figure 5: Robust stopping problem ($|\mathcal{S}| = 200$, $N = 20$, $m = 80$)

to the confidence level ω in the nominal chain. Clearly $V_0^{nr} \geq V_{0,ml}^{r,\omega}$. We will call the ratio

$$M(\omega) = \frac{\sum_{s \in \mathcal{S}} V_{0,ml}^{r,\omega}(s)}{\sum_{s \in \mathcal{S}} V_0^{nr}(s)},$$

the relative *nominal* performance of the robust policy. The ratio $M(\omega)$ measures the loss associated with using a robust policy designed for a confidence level ω. Clearly $M(\omega) \leq 1$ and we expect the ratio to decrease as ω increases.

The second performance measure we consider is the worst case performance of the non-robust policy. Define $V_{N,w}^{nr} = h$. For $t = 0, \ldots, N-1$, define

$$V_{t,w}^{nr}(s) = \max\Big\{g_t(s), f_t(s) + \inf_{p \in \mathcal{P}(s,\omega)} \mathbf{E}^p[V_{t+1,w}^{nr}]\Big\},$$

where $\mathcal{P}(s,\omega)$ is the set of conditional measures for state $s \in \mathcal{S}$ corresponding to a confidence level ω. Thus, $V_{0,w}^{nr}$ denotes the worst case value of the non-robust policy. We will call the ratio

$$R(\omega) = \frac{\sum_{s \in \mathcal{S}} V_0^{r,\omega}(s)}{\sum_{s \in \mathcal{S}} V_{0,w}^{nr}(s)},$$

the relative *worst-case* performance of the robust policy. Since the robust policy optimizes the worst case value, $R(\omega) \geq 1$ and we expect the ratio to increase as ω increases. The ratio $R(\omega)$ measures the relative gain associated with using the robust policy when the transition probabilities are ambiguous.

Figures 3-5 plot the relative nominal and worst case performance for three different simulation runs. The plots show that the relative loss in nominal performance of the robust policy even at $\omega = 0.95$ is approximately 15%. On the other hand the worst case performance improves with ω and is greater than 250% at $\omega = 0.95$. This behavior may be explained by the fact the robust and non-robust optimal policies differ only on a few states in the entire trellis. Thus, the robust policy appears to be able to track the mean behavior while at the same time improve the worst case behavior by altering the action in a small number of critical states. The relative nominal performance appears to be fairly stable as a function of the time horizon N, the size of the state space $|\mathcal{S}|$ and the number of 1-step neighbors m. These numerical experiments are clearly quite preliminary. We are currently conducting experiments to further understand the relative merits of the robust approach.

5.2 Robust infinite horizon dynamic programs

In this section we contrast the computational effort required to solve discounted infinite horizon robust and non-robust DPs. This comparison is done by averaging the CPU time and the number of iterations required to solve randomly generated problems. The details of the experiments are given below. All computations were done in the MATLAB6.1 R12 computing environment and, therefore, only the relative values of the run times are significant.

The first set of experiments compared the required computational effort as a function of the uncertainty level. For this set of experiments, the size $|\mathcal{S}|$ of the state space was set to $|\mathcal{S}| = 500$, the number of actions $|\mathcal{A}(s)|$ was $|\mathcal{A}(s)| = 10$, and the discount rate $\lambda = 0.9$. The rewards $r(s, a, s')$ were assumed to be independent of the future state s' and distributed uniformly over $[0, 10]$. The state transition was also randomly generated. The ambiguity in the transition structure was assumed to be given by $\mathcal{L}_2$ approximation to the relative entropy sets (see (46)).

For each value of $\omega = (0.05, 0.1, \ldots, 0.95)$ the results were averages over $N = 10$ random instances. The random instances were solved using both value iteration and policy iteration. The robust value iteration followed the algorithm described in Figure 1 and was terminated once the difference between successive iterates was less than $\tau = 10^{-6}$. However, the robust policy iteration did not entirely follow the algorithm in Figure 2 – instead of solving (31), the value function of the policy π_n was computed by iteratively solving the operator equation (29).

The results for this set of experiments are shown in Table 1. The columns labeled *iter* display the number of iterations and the columns marked *time* display the run time in seconds. From these results it is clear that both the run times and the number of iterations is insensitive to ω. Both the non-robust and robust DP require approximately the same number of iterations to solve the problem. However, the run time per robust value iteration is close to twice that of the run time per non-robust value iteration.

The second set of experiments compared the run time as a function of size $|\mathcal{S}|$ of the state space. In this set of experiments, the uncertainty level $\epsilon = 0.95$ and the discount rate $\lambda = 0.9$. For each value of $|\mathcal{S}| = 200, 400, \ldots, 1000$, the results were averaged over $N = 10$ random instances. Each random instance was generated just as in the first set of experiments. As before, each instance was solved using both value iteration and policy iteration.

The results for this set of experiments in shown in Table 2. From the results in the table, it appears that the number of iterations is insensitive to the size $|\mathcal{S}|$. Moreover, the robust and non-robust DP algorithms require approximately the same number of iterations. Therefore, based on Lemma 5, we expect that the

	Value iteration				Policy iteration			
	Non-robust		Robust		Non-robust		Robust	
ϵ	iter	time (sec)	iter	time (sec)	iter	time (sec)	iter	time (sec)
0.05	153.7	4.73	152.9	8.23	3.9	4.29	3.0	6.67
0.10	153.6	4.66	152.7	8.48	3.9	4.31	3.0	6.87
0.15	153.9	4.69	152.6	8.60	3.6	4.36	2.9	6.79
0.20	153.8	4.61	152.2	8.64	3.5	4.21	2.8	6.60
0.25	154.0	4.80	152.1	8.50	3.9	4.32	2.8	6.54
0.30	154.2	4.72	151.9	8.73	3.8	4.21	2.8	6.73
0.35	153.7	4.68	151.8	8.71	3.6	4.28	2.9	6.88
0.40	154.1	4.70	151.8	8.79	3.8	4.26	2.8	6.72
0.45	153.4	4.67	151.8	8.84	3.7	4.37	2.8	6.76
0.50	154.2	4.69	151.8	8.69	3.3	4.20	2.8	6.70
0.55	153.5	4.65	151.6	8.84	3.3	4.35	2.8	6.74
0.60	153.5	4.72	151.5	8.84	3.3	4.39	2.8	6.76
0.65	154.1	4.75	151.5	8.88	3.5	4.40	2.9	6.96
0.70	153.9	4.71	151.5	8.83	3.7	4.36	2.9	6.90
0.75	153.3	4.73	151.4	8.80	3.3	4.29	2.9	6.90
0.80	153.2	4.64	151.1	8.86	3.8	4.30	3.0	7.14
0.85	154.1	4.68	151.0	8.83	3.6	4.24	3.0	7.10
0.90	153.4	4.76	150.8	8.79	3.4	4.33	3.0	7.10
0.95	153.5	4.74	150.7	8.84	3.7	4.26	3.0	7.12

Table 1: Robust DP vs ϵ $\big(|\mathcal{S}| = 500,\ |\mathcal{A}(s)| = 10,\ \lambda = 0.9\big)$

run time of robust DP is at most a logarithmic factor higher than the run time of the non-robust version. Regressing the run time t_v of the non-robust value iteration on the sample size $|\mathcal{S}|$, we get

$$\log(t_v) \approx 2.1832 \log(|\mathcal{S}|) - 11.6401$$

Regressing the run time t_p of the non-robust policy iteration on the sample size $|\mathcal{S}|$, we get that

$$\log(t_p) \approx 2.0626 \log(|\mathcal{S}|) - 11.7612$$

The regression results are plotted in Figure 6. The upper plot corresponds to value iteration and the bottom plot corresponds to policy iteration. The dotted line in both plots is the best fit line obtained by regression and the solid line is the observed run times. Clearly the regression approximation fits the observed run times very well.

In the upper plot of Figure 7, the solid line corresponds to $\log(t_{rv})$, where t_{rv} is the observed run time of robust value iteration, and the dotted line corresponds to the upper bound $\log(\bar{t}_{rv})$ expected on the basis of Lemma 5, i.e.

$$\begin{aligned} \log(\bar{t}_{rv}) &= \log\log(|\mathcal{S}|) + \log(t_v), \\ &= \log\log(|\mathcal{S}|) + 2.1832 \log(|\mathcal{S}|) - 11.6401 \end{aligned}$$

Clearly, the expected upper bound dominates over the observed run times, i.e $t_{rv} << t_v \log(|\mathcal{S}|)$. In the second plot of Figure 7 we plot $\log(t_{rp})$, where t_{rp} is the run time of robust policy iteration, and the corresponding expected upper bound $\log(\bar{t}_{rp})$. Once again, the upper bound clearly dominates.

	Value iteration				Policy iteration			
	Non-robust		Robust		Non-robust		Robust	
ϵ	iter	time (sec)	iter	time (sec)	iter	time (sec)	iter	time (sec)
200	154.2	1.000	153.6	0.710	3.4	0.480	3.5	0.329
400	154.3	3.840	153.7	5.350	3.3	1.720	3.4	2.421
600	153.6	9.420	153.4	17.654	3.1	3.960	3.4	7.939
800	154.8	19.180	153.6	42.129	3.2	7.320	3.6	19.939
1000	154.9	33.320	153.5	83.986	3.2	11.460	3.2	36.368
1200	153.8	49.480	153.2	143.217	3.5	17.860	3.7	68.953
1400	155.2	66.300	153.2	223.425	3.1	22.020	3.5	104.234
1600	154.2	87.040	153.6	337.874	3.6	33.680	3.5	157.546
1800	153.8	112.160	153.4	483.529	3.2	40.600	3.3	214.227
2000	154.2	136.360	153.7	649.510	3.5	55.340	3.8	322.526

Table 2: Robust DP vs $|\mathcal{S}|$ $\big(|\mathcal{A}(s)| = 10,\ \epsilon = 0.95,\ \lambda = 0.9\big)$

These computational results are still preliminary and there are many unresolved issues. For example, although the bounds dominate the run times of robust DP, the two lines appear to converge leading one to believe that the bound may not hold for larger state spaces. However, recall that the bounds are constructed using linear regression and, therefore, there is the possibility that the bound will shift upward when larger state spaces are considered. The codes for both non-robust and robust DP needs to be optimized before one can completely trust the run times and iterations.

6 Conclusion

In this paper we propose a robust formulation for the discrete time DP. This formulation attempts to mitigate the impact of errors in estimating the transition probabilities by choosing a maximin optimal policy, where the minimization is over a set of transition probabilities. This set summarizes the limited knowledge that the decision maker has around the transition probabilities of the underlying Markov chain. A natural family of sets describing the knowledge of the decision maker are the confidence regions about the maximum likelihood estimates of the transition probability. This family of sets was first introduced in Nilim and El Ghaoui (2002). Since these confidence regions are described in terms of the relative entropy or the Kullback-Liebler distance, we are led to the sets described in Section 4.1. The family of relative entropy based sets can be easily parameterized by setting the desired confidence level. We also introduce two other families of sets that are approximations of the relative entropy based sets.

Since the transition probabilities are ambiguous, every policy now has a set of measures associated with it. We prove that when this set of measures satisfies a certain Rectangularity property most of important results in DP theory, such as the Bellman recursion, the optimality of deterministic Markov policies, the contraction property of the value iteration operator, etc., extend to natural robust counterparts. On the computational front, we show that the the computational effort required to solve the robust DP corresponding to sets of conditional measures based on confidence regions is only modestly higher than that required to solve the non-robust DP. Our preliminary computational results appear to confirm this experimentally. While parts of the theory presented in this paper have been addressed by other authors, we provide a unifying framework for the theory of robust DP.

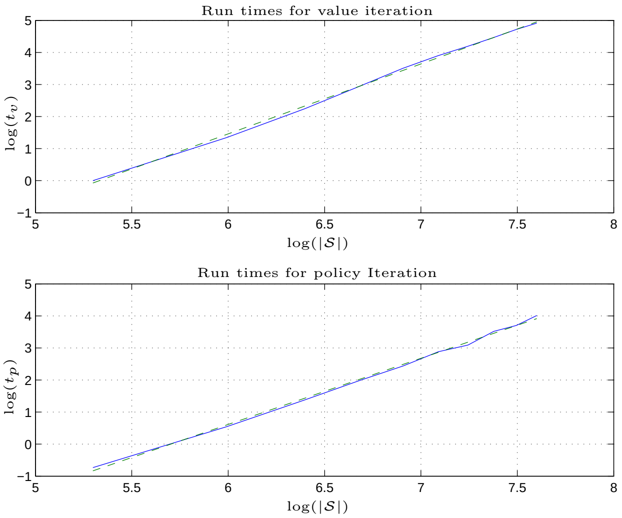

Figure 6: Regression results for non-robust DP run times

The robust value function V^* provides a lower bound on the achievable performance; one can also define an optimistic value function $\bar{V}^*$ that provides an upper bound on the achievable performance. All the results in this paper imply corresponding results for the optimistic value function, i.e. in particular there is value iteration and a policy iteration algorithm that efficiently characterizes the optimistic value function.

Some unresolved issues that remain are as follows. The computational results presented in this paper are very preliminary. While the initial results are promising, more experiments need to be performed in order to better understand the performance of robust DP on practical examples. As indicated in the introduction, we restricted our attention to problems where the non-robust DP is tractable. In most of the interesting applications of DP, this is not the case and one has to resort to approximate DP. One would, therefore, be interested in developing the robust counterpart of approximate DP. Such an approach might be able to prevent instabilities observed in approximate DP (Bertsekas and Tsitsiklis, 1996).

Acknowledgments

The authors would like to thank the anonymous referees for insightful comments that have substantially improved the presentation and the conceptual content of the paper.

A Consequences of Rectangularity

We will begin with an example that illustrates the inappropriateness of the Rectangularity in a finite horizon setting. This example is a dynamic version of the Ellsberg Urn problem (Ellsberg, 1961) discussed in Epstein and Schneider (2001).

Suppose an urn contains 30 red balls and 60 balls that are either blue or green. At time 0 a ball is drawn from the urn and the the color of the ball is revealed at time $t = 2$. At the intermediate time $t = 1$ the decision maker is told whether the drawn ball is green. Thus, the state transition structure is as shown in Figure 8 where $p_b = \mathbf{P}\{\text{ball is blue}\}$.

Suppose $p_b \in [\underline{p}, \overline{p}] \subseteq [0, 2/3]$ is ambiguous. Consider the robust optimal stopping problem where the

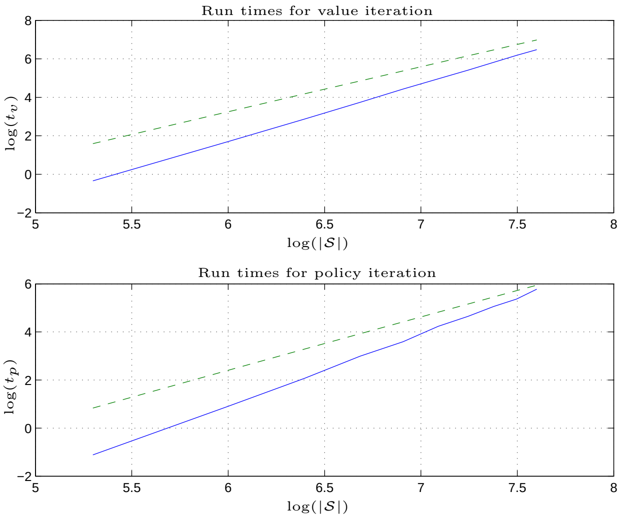

Figure 7: Run times of robust DP and corresponding bounds

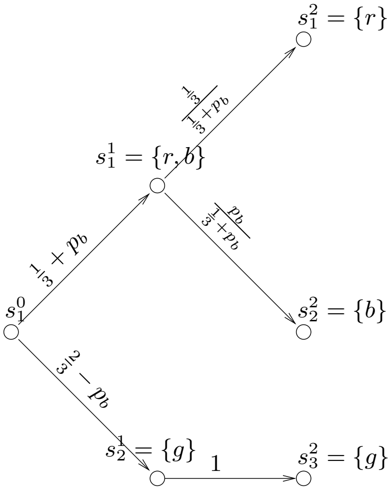

Figure 8: Dynamic Ellsberg experiment

state transition is given by Figure 8. In each state $s \in \mathcal{S}_t$ at time $t = 0, 1$ there are are two actions $\{s, c\}$ available, where c denotes *continue* and s denotes *stop*. Let $\bar{\pi} = (\bar{d}_0, \bar{d}_1)$ denote the policy that chooses the deterministic action c in every state $s \in \mathcal{S}_t$, $t = 0.1$. Then the state-transition structure in Figure 8 implies that the conditional measures consistent with the decision rules $\bar{d}_i$, $i = 0, 1$ are given by

$$\begin{aligned}
\mathcal{T}^{\bar{d}_0} &= \left\{\big(p(s_1^1 \mid s_1^0), p(s_2^1 \mid s_1^0)\big) = (1/3 + \alpha, 2/3 - \alpha) : \alpha \in [\underline{p}, \overline{p}]\right\}, \\
\mathcal{T}^{\bar{d}_1} &= \left\{\big(p(s_1^2 \mid s_1^1), p(s_2^2 \mid s_1^1)\big) = \left(\frac{1/3}{1/3 + \alpha}, \frac{\alpha}{1/3 + \alpha}\right), p(s_2^2 \mid s_2^1) = 1 : \alpha \in [\underline{p}, \overline{p}]\right\}.
\end{aligned}$$

Thus,

$$\mathcal{T}^{\bar{d}_0} \times \mathcal{T}^{\bar{d}_1} = \left\{ \begin{array}{l} \big(p(s_1^1 \mid s_1^0), p(s_2^1 \mid s_1^0)\big) = (1/3 + \alpha, 2/3 - \alpha), \\ \big(p(s_1^2 \mid s_1^1), p(s_2^2 \mid s_1^1)\big) = \left(\frac{1/3}{1/3+\alpha'}, \frac{\alpha'}{1/3+\alpha'}\right), p(s_2^2 \mid s_2^1) = 1 \end{array} : \alpha, \alpha' \in [\underline{p}, \overline{p}] \right\},$$

where α and α' need not be equal. However, the set of measures $\mathcal{T}^{\bar{\pi}}$ consistent with the policy $\bar{\pi}$ satisfies

$$\begin{aligned} \mathcal{T}^{\bar{\pi}} &= \left\{ \begin{array}{l} \big(p(s_1^1 \mid s_1^0), p(s_2^1 \mid s_1^0)\big) = (1/3+\alpha, 2/3-\alpha)\,, \\ \big(p(s_1^2 \mid s_1^1), p(s_2^2 \mid s_1^1)\big) = \left(\frac{1/3}{1/3+\alpha}, \frac{\alpha}{1/3+\alpha}\right), p(s_2^2 \mid s_2^1) = 1 \end{array} : \alpha \in [\underline{p}, \overline{p}] \right\}, \\ &\neq \mathcal{T}^{\bar{d}_0} \times \mathcal{T}^{\bar{d}_1}. \end{aligned}$$

The problem arises because the information structure in Figure 8 assumes that there is a single urn that decides that conditional measures at both epochs $t = 0, 1$; whereas, Rectangularity demands that the conditional measures at epochs $t = 0, 1$ be independent, i.e. in this case, they should be determined by an *independent* copy of the urn used at $t = 0$.

Assuming that Rectangularity holds in this setting is equivalent to assuming that apriori distribution on the composition of the urn is given by

$$(p_r, p_b, p_g) \in \mathcal{P} = \left\{ \frac{1}{3}\left(\frac{1/3+\alpha}{1/3+\alpha'}\right), \alpha'\left(\frac{1/3+\alpha}{1/3+\alpha'}\right), \frac{2}{3} - \alpha \right\}.$$

A very counterintuitive prior indeed ! This example clearly shows that Rectangularity may not always be an appropriate property to impose on an AMDP. Inspite of the counterexample above, Rectangularity is often appropriate for finite horizon AMDPs because the sources of the ambiguity in different periods are typically independent of each other.

Rectangularity implies that the adversary is able to choose a different conditional measure every time a state-action pair (s, a) is encountered. This adversary model should not raise an alarm in a finite horizon setting where a state-action pair is never revisited. However, the situation is very different in a infinite horizon setting where a state-action can be revisited. In this setting the Rectangularity may not be appropriate situations where there is ambiguity but the transition probabilities are not dynamically changing. Deciding whether Rectangularity is appropriate can often be a function of the time scale of events. Suppose one is interested in a robust analysis of network routing algorithms where the action in each node is the choice of the outgoing edge and the ambiguity is with respect to the delay on the network edges. For a traffic network the Rectangularity assumption might be appropriate because the time elapsed in returning to a node is sufficiently long so that the parameters could have shifted. On the other hand, for data networks that operate at much higher speeds the ambiguity might be evolve on a slower time scale, and therefore, Rectangularity might not be appropriate. On a positive note, Lemma 3 shows that the problems with Rectangularity disappear if one restricts the decision maker to stationary policies.

References

Bagnell, J., Ng, A., and Schneider, J. (2001). Solving uncertain Markov decision problems. Technical report, Robotics Inst., CMU.

Ben-Tal, A. and Nemirovski, A. (1997). Robust truss topology design via semidefinite programming. *SIAM J. Optim.*, 7(4):991–1016.

Ben-Tal, A. and Nemirovski, A. (1998). Robust convex optimization. *Math. Oper. Res.*, 23(4):769–805.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming.* Athena Scientific.

Cover, T. M. (1991). Universal portfolios. *Mathematical Finance*, 1(1):1–29.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* John Wiley & Sons, New York.

de Farias, D. and Van Roy, B. (2002). The linear programming approach to approximate dynamic programming. Submitted to *Oper. Res.*

Ellsberg, D. (1961). Risk, ambiguity and the Savage axioms. *Quart. J. Econ.*, 25(643-669).

Epstein, L. G. and Schneider, M. (2001). Recursive multiple priors. Technical Report 485, Rochester Center for Economic Research. Available at http://rcer.econ.rochester.edu. To appear in *J. Econ. Theory.*

Epstein, L. G. and Schneider, M. (2002). Learning under Ambiguity. Technical Report 497, Rochester Center for Economic Research. Available at http://rcer.econ.rochester.edu.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique priors. *J. Math. Econ.*, 18:141–153.

Goldfarb, D. and Iyengar, G. (2003). Robust portfolio selection problems. *Math. Oper. Res.*, 28(1):1–38.

Hansen, L. P. and Sargent, T. J. (2001). Robust control and model uncertainty. *American Economic Review*, 91:60–66.

Littman, M. (1994). memoryless policies: Theoretical limitations and practical results. In Cliff, D., Husbands, P., and Wilson, S. W., editors, *From Animals to Animats: SAB '94*, pages 238–245. MIT Press.

Nilim, A. and El Ghaoui, L. (2002). Robust Solutions to Markov Decision Problems with Uncertain Transition Matrices. Submitted to *Operations Research*. UC Berkeley Tech Report UCB-ERL-M02/31.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley series in probability and mathematical statistics. John Wiley & Sons.

Satia, J. K. (1968). *Markovian Decision Process with Uncertain Transition Matrices or/and Probabilistic Observation of States.* PhD thesis, Stanford University.

Satia, J. K. and Lave, R. L. (1973). Markov Decision Processes with Uncertain Transition Probabilities. *Oper. Res.*, 21(3):728–740.

Shapiro, A. and Kleywegt, A. J. (2002). Minimax analysis of stochastic problems. To appear in *Optimization Methods and Software.*

Tsitsiklis, J., Simester, D., and Sun, P. (2002). Dynamic optimization for direct marketing problem. Presented at INFORMS 2002.

White, C. C. and Eldieb, H. K. (1994). Markov decision processes with imprecise transition probabilities. *Oper. Res.*, 43:739–749.