Mathematics of Operations Research

Robust Dynamic Programming

Garud N. Iyengarhttp://www.columbia.edu/~gi10,

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 30, No. 2, May 2005, pp. 257–280
ISSN 0364-765X | EISSN 1526-5471 | 05 | 3002 | 0257

DOI 10.1287/moor.1040.0129

Robust Dynamic Programming

Garud N. Iyengar
IEOR Department, Columbia University, New York, New York 10027,
garud@ieor.columbia.edu, http://www.columbia.edu/∼gi10

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 of the main results for finite and infinite horizon DP extend to natural robust counterparts. We discuss techniques from Nilim and El Ghaoui [17] for constructing suitable sets of conditional measures that allow one to efficiently solve for the optimal robust policy. We also show that robust DP is equivalent to stochastic zero-sum games with perfect information.

Key words: dynamic programming; robust optimization; Markov decision processes; ambiguity
MSC2000 subject classification: Primary: 90C39, 90C47; secondary: 90C40, 90C25
OR/MS subject classification: Primary: Dynamic programming/optimal control, decision analysis-risk; secondary: probability-Markov processes
History: Received December 17, 2002; revised October 10, 2003, and June 11, 2004.

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 [20] 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.

The DP formalism encodes information in the form of a "reward-to-go" function (see Puterman [20] 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 [5], de Farias and Van Roy [8]).

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 [17], Tsitskilis et al. [24]). Degradation in performance due to estimation errors in parameters has also been observed in other contexts (Ben-Tal and Nemirovski [3], Goldfarb and Iyengar [14]). 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 $\mathscr{P}(s, a)$ 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 [10]). 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 [10]), 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 [22], White and Eldieb [25] and Bagnell et al. [2] 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 prespecified 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 [23] 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 [12], Epstein and Schneider [10, 11], Hansen and Sargent [15]). 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.

While this paper was being prepared for submission, we became aware of a technical report by Nilim and El Ghaoui [17] where they independently formulate robust DP and develop a robust Bellman recursion. Although not explicitly stated, the rectangularity assumption is implicit in their construction. They identify sets of conditional measures based on likelihood measures that have the following desirable properties: The 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; and the complexity of solving the robust DP corresponding to these families of sets is only modestly larger than the nonrobust counterpart.

This paper is organized as follows. In §2 we formulate finite horizon robust DP and the "rectangularity" property that leads to the robust counterpart of the Bellman recursion. In §3 we formulate the robust extension of discounted infinite horizon DP. In §4 we describe three families of sets of conditional measures that are all based on the relative entropy or the

Kullback-Leibler distance. The results in this section, although independently obtained, are not new and were first obtained by Nilim and El Ghaoui [17] (see Nilim and El Ghaoui [18]). In §5 we show that the robust DP is equivalent to stochastic two-player zero-sum games with perfect information. Section 6 concludes the paper with some 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 nondiscrete 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 [10]).

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 only a function of the state $s \in \mathcal{S}_N$.

We will refer to the collection of objects $\{T, \{\mathcal{S}_t, \mathcal{A}_t, \mathcal{P}_t, r_t(\cdot, \cdot, \cdot)\colon t \in T\}\}$ as a finite horizon *ambiguous Markov decision process* (AMDP). The notation above is a modification of that in Puterman [20] and the structure of ambiguity is motivated by Epstein and Schneider [10].

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 $d_t\colon \mathcal{H}_t \mapsto \mathcal{M}(\mathcal{A}(s_t))$. 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

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

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

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

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., $\pi = (d_t\colon t \in T)$. Given the ambiguity in the conditional measures, a policy π induces a collection of measure on the history space $\mathscr{H}_N$. We assume that the set $\mathscr{T}^\pi$ of measures consistent with a policy π has the following structure.

Assumption 2.1 (Rectangularity). *The set $\mathscr{T}^\pi$ of measures consistent with a policy π is given by*

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

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

The rectangularity assumption is motivated by the structure of the recursive multiple priors in Epstein and Schneider [10]. We will defer discussing the implications of 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.

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

where $\mathbf{E}^{\mathbf{P}}$ denotes the expectation with respect to the fixed measure $\mathbf{P} \in \mathscr{T}^\pi$. 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 [4]). Let Π denote the set of all history dependent policies. Then the goal of *robust* DP is to characterize the *robust* value function

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

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 be interpreted in an adversarial setting: The decision maker chooses π; an adversary observes π, and chooses a measure $\mathbf{P} \in \mathscr{T}^\pi$ that minimizes the reward. In this context, rectangularity is a form of an independence assumption: The choice of particular distribution $\bar{p} \in \mathscr{P}(s_t, a_t)$ 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 that is crucial for establishing the robust counterpart of the Bellman recursion (see Theorem 2.1). Similar separability assumptions also appear in the robust control literature in the context of obtaining lower bounds for system performance (see, e.g., Nilim and El Ghaoui [17]). The rectangularity assumption is not always appropriate (see Appendix A for an example of such a situation). This assumption can be somewhat justified in the context of finite-horizon problems by invoking time-inhomogeneity. However, such an explanation does not extend to infinite horizon models. We will discuss the implications of rectangularity in infinite horizon models in §3.

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

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

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

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

Remark 2.1. 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 $n, n+1, \dots, N-1$, starting from the history h_n; i.e.,

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

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

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

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

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

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

Theorem 2.1 (Bellman Equation). *The set of functions* $\{V_n^*\colon n = 0, 1, \dots, N\}$ *satisfies the following robust Bellman equation*:

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

Proof. From (9) it follows that

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

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:

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

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 to the (randomized) decision rule d_n. Then $V^{\pi}_{n+1}(h_n, d_n(h_n)(\omega), s_{n+1}(\omega)) \leq V^*_{n+1}(h_n, d_n(h_n)(\omega), s_{n+1}(\omega))$. 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 $V^*_{n+1}(h_{n+1}) = \sup_{\pi\in\Pi_{n+1}}\{V^{\pi}_{n+1}(h_{n+1})\}$, it follows that for all $\epsilon > 0$ there exists a policy $\pi^{\epsilon}_{n+1} \in \Pi_{n+1}$ such that $V^{\pi^{\epsilon}_{n+1}}_{n+1}(h_{n+1}) \geq V^*_{n+1}(h_{n+1}) - \epsilon$, for all $h_{n+1} \in \mathcal{H}_{n+1}$. For all $d_n \in \mathcal{D}_n$, $(d_n, \pi^{\epsilon}_{n+1}) \in \Pi_n$. Therefore,

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

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.

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

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. □

As mentioned before, while this paper was being prepared for publication we became aware of a technical report by Nilim and El Ghaoui [17] where they independently formulate robust DP and present a robust Bellman recursion.

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

Corollary 2.1. *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* $n = 0, \ldots, N-1$,

$$V^*_n(h_n) = \sup_{\pi\in\Pi_D}\{V^{\pi}_n(h_n)\}.$$

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 2.2 (Markov Optimality). *For all $n = 0, \dots, N$, 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*

$$V_n^*(s_n) = \sup_{a \in \mathscr{A}(s_n)} \left\{ \inf_{p \in \mathscr{P}_n(s_n, a)} \mathbf{E}^p\big[r_n(s_n, a, s) + V_{n+1}^*(s)\big]\right\}, \quad n \in T. \tag{16}$$

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

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

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 $s' \in \mathscr{S}$, computing $V_n^*(s)$ reduces to a collection of optimization problems. Suppose the action set $\mathscr{A}(s)$ is finite. Then the optimal decision rule d_n^* at epoch n is given by

$$d_n^*(s) = \arg\max_{a \in \mathscr{A}(s)} \left\{ \inf_{p \in \mathscr{P}_n(s, a)} \mathbf{E}^p\big[r_n(s, a, s') + V_{n+1}(s')\big]\right\}.$$

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

As noted in Remark 1, Theorem 2.2 implies the following result for the optimistic value function $\bar{V}_n^*$.

Theorem 2.3. *For $n = 0, \dots, N$, the optimistic value function $\bar{V}_n^*(h_n)$ is a function of the current state s_n alone, and*

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

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

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

3. Infinite horizon robust dynamic programming. In this section we formulate 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 [21] and Satia and Lave [22]. A special case of the robust DP where the decision maker is restricted to stationary policies appears in Bagnell et al. [2]. 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 §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 $t \in T$; i.e., time is homogeneous. Thus, one is forced to carefully delineate the dynamic structure of uncertainty. We consider two distinct models for the uncertainty, or equivalently, the adversary.

(i) Static model: 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.

(ii) Dynamic model: 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. Thus, in this model, the set $\mathcal{T}^\pi$ of measures consistent with a policy π satisfies rectangularity; i.e., $\mathcal{T}^\pi = \prod_{t \in T} \mathcal{T}^{d_t}$.

As mentioned in the introduction, the goal of the robust formulation is to systematically mitigate the effect of errors associated with estimating the state transitions; i.e., the state transition is, in fact, fixed but the decision maker is only able to estimate it to within a set. Thus, the static model is appropriate for this scenario. However, computing the optimal policy for the static model is NP-hard; therefore, we will restrict attention to the dynamic model. Clearly the value function in the dynamic model is a lower bound for the value function in the static model. We contrast the implications of the two models in Lemma 3.3. Bagnell et al. [2] 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

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

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

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

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

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

As noted in Remark 2.1, all the results in this section imply a corresponding result for the optimistic value function $\overline{V}_\lambda^*$ 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 2.2.

Theorem 3.1 (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 [20]. □

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 $V\in\mathbf{V}$,

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

Theorem 3.2 (Bellman Equation). *The operator $\mathcal{L}_{\mathcal{D}}$ satisfies the following properties*:

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

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

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

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

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

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

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

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

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

Then

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

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

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

i.e., $\|\mathcal{L}U - \mathcal{L}V\| \le \lambda\|U - V\| + 2\epsilon$. 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 $d_t^\pi \in \mathscr{D}$, for all $t \geq 0$. Then

$$V(s) = \mathscr{L}_{\mathscr{D}} V(s),$$

$$\geq \inf_{p_0 \in \mathscr{P}(s, d_0^\pi(s))} \mathbf{E}^{p_0}[r(s, d_0^\pi(s), s_1) + \lambda V(s_1)] \tag{24}$$

$$\geq \inf_{p_0 \in \mathscr{P}(s, d_0^\pi(s))} \mathbf{E}^{p_0}\Big[r(s, d_0^\pi(s), s_1) + \lambda \inf_{p_1 \in \mathscr{P}(s_1, d_1^\pi(s_1))} \mathbf{E}^{p_1}[r(s_1, d_1^\pi(s_1), s_2) + \lambda V(s_2)]\Big] \tag{25}$$

$$= \inf_{\mathbf{P} \in \mathscr{T}^\pi} \mathbf{E}^{\mathbf{P}}\bigg[\sum_{t=0}^{1} r(s_t, d_t^\pi(s_t), s_{t+1}) + \lambda^2 V(s_{t+1})\bigg], \tag{26}$$

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 $n \geq 0$,

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

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

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

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

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

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 $V(s) = \sup_{\{\pi: d_t^\pi \in \mathscr{D}, \forall t\}}\{V^\pi(s)\}$. □

Corollary 3.1. *The properties of the operator $\mathscr{L}_{\mathscr{D}}$ 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*

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

(b) *The value function V_λ^* is the unique solution of the operator equation*

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

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 $\mathscr{D} = \{d\}$ and $\mathscr{D} = \prod_{s \in \mathscr{S}} \mathscr{A}(s)$ respectively. □

Theorem 3.1 and part (b) of Corollary 3.1 for the special case of finite state and action spaces appears in Satia [21] and Satia and Lave [22] with an additional assumption that the set of conditional measures $\mathscr{P}(s, a)$ is convex. (Their proof, in fact, extends to nonconvex

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

Figure 1. Robust value iteration algorithm.

$\mathscr{P}(s,a)$.) Also, they do not explicitly prove that the solution of (30) is indeed the robust value function. Theorem 3.2 for general $\mathscr{D}$, and in particular for $\mathscr{D} = \{d\}$, is new. The special case $\mathscr{D} = \{d\}$ is crucial for establishing the policy improvement algorithm.

From Theorem 3.2, Corollary 3.1, 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 nonrobust DPs (see §6.3.2 in Puterman [20] for details). The following lemma establishes this approximation result for the robust value iteration algorithm.

Lemma 3.1. *Let $\widetilde{V}$ be the output of the robust value iteration algorithm shown in Figure 1. Then*

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

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

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

Then, the policy $\pi = (d, d, \ldots)$ is ϵ-optimal. Moreover, the overall complexity of computing the policy π is $\mathscr{O}(C|S|(\log(R/\epsilon))/(\log(1/\lambda)))$, where C is cost of computing $\sup_{a \in \mathscr{A}(s)} \{\inf_{p \in \mathscr{P}(s,a)} \mathbf{E}^p[r(r,a,s') + \lambda V(s')]\}$ for a fixed $s \in \mathscr{S}$.

Proof. Since Lemma 3.2 establishes that $\mathscr{L}_D$ is a contraction operator, this result is a simple extension of Theorem 6.3.1 in Puterman [20]. The details are left to the reader. □

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

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

For this iteration to be efficient, one must be able to efficiently solve the optimization problem $\inf_{p \in \mathscr{P}(s,a)} \mathbf{E}^p[v]$ for a specified $s \in \mathscr{S}$, $a \in \mathscr{A}(s)$, and $v \in \mathbf{R}^{|\mathscr{S}|}$. These optimization problems are identical to those solved in finite state problems. In §4 we show that for suitable choices for the set $\mathscr{P}(s,a)$ 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 3.2 below establishes that policy evaluation is a robust optimization problem.

Lemma 3.2 (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*

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

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

PROOF. The constraint in (31) can be restated as $V \leq \mathscr{L}_d V$, where $\mathscr{L}_d = \mathscr{L}_{\mathscr{D}}$ with $\mathscr{D} = \{d\}$. Corollary 3.1 implies that $V^\pi = \mathscr{L}_d V^\pi$; i.e., V^π is feasible for (31). Therefore, the optimal value of (31) is at least $\sum_{s\in\mathscr{S}} \alpha(s)V^\pi(s)$.

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

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

Then for any V feasible for (31),

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

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 $\alpha(s) > 0$, $s \in \mathscr{S}$, it follows that the value of (31) is at most $\sum_{s\in\mathscr{S}} \alpha(s)V^\pi(s)$. This establishes the result. □

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 $\mathscr{S}$ is finite and the robust constraint can be reformulated as a small collection of deterministic constraints. In §4 we discuss some natural candidates for the set $\mathscr{P}(s, a)$ 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 an m-step look-ahead value iteration (Puterman [20]).

Lemma 3.1 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 [20]. A rudimentary version of robust policy iteration algorithm for this special case appears in Satia and Lave [22] (see also Satia [21]). 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 $s \in \mathscr{S}$, fix $p_s \in \mathscr{P}(s, d(s))$. Solve the set of equations

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

Since $\lambda < 1$, this set of equations has a unique solution. See Theorem 6.1.1 in Puterman [20].

```
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 $\tilde{V} \leftarrow \mathscr{L}_{\mathscr{D}} V^{\pi_n}$, $\mathscr{D} = \prod_{s\in\mathscr{S}} \mathscr{A}(s)$
  For each $s \in \mathscr{S}$, choose
      $d_{n+1}(s) \in \Big\{a \in \mathscr{A}(s)\colon \inf_{p\in\mathscr{P}(s,a)} \mathbf{E}^p[r(s, a, s') + \lambda V(s')] \geq \tilde{V}(s) - \epsilon\Big\}$;
    setting $d_{n+1}(s) = d_n(s)$ if possible.
  while $(d_{n+1} \neq d_n)$ do
    $n = n + 1$; Solve (31) to compute $V^{\pi_n}$. Set $\tilde{V} \leftarrow \mathscr{L}_{\mathscr{D}} V^{\pi_n}$, $\mathscr{D} = \prod_{s\in\mathscr{S}} \mathscr{A}(s)$
    For each $s \in \mathscr{S}$, choose
      $d_{n+1}(s) \in \Big\{a \in \mathscr{A}(s)\colon \inf_{p\in\mathscr{P}(s,a)} \mathbf{E}^p[r(s, a, s') + \lambda V(s')] \geq \tilde{V}(s) - \epsilon\Big\}$;
    setting $d_{n+1}(s) = d_n(s)$ if possible.
  end while
return $d_{n+1}$
```

FIGURE 2. Robust policy iteration algorithm.

(b) Fix V, and solve

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

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

However, it is not clear, and Satia and Lave [22] 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 [20]):

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

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 $\mathscr{T}^{\pi}$ 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 $p_{sa} \in \mathscr{P}(s, a)$ 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 3.3 (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 $\bar{p}\colon \mathscr{S} \mapsto \mathscr{M}(\mathscr{S})$ such that $\bar{\mathbf{p}}_s \in \mathscr{P}(s, d(s))$ for all $s \in \mathscr{S}$, 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 nonrobust 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 nonrobust 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,

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

Iterating this bound for n time steps, we get

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

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 $p_{sa} \in \mathscr{P}(s, a)$ 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. [2] establish that when the set $\mathscr{P}(s, a)$ of conditional measures is convex and the decision maker is restricted to stationary policies, the optimal policies for the decision maker and the adversary are the same in both the static and dynamic models. We extend this result to nonconvex 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 [16]), the problem is tractable provided the sets $\mathcal{P}(s, a)$ are "nice" convex sets. In particular, the problem is tractable for the families of sets discussed in §4.

Lemma 3.3 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 [6]).

4. Computational complexity.

Sections 2 and 3 were devoted to extending results from nonrobust DP theory. In this 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. Although independently obtained, most of the results in this section are not new; they first appeared in Nilim and El Ghaoui [17] (see also Nilim and El Ghaoui [18]). We include them here for completeness.

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 (s, a) is given by $\inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p[v]$. In this section, we discuss 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. 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. Finally, we discuss the issue of error propagation in the Bellman equation when the inner optimization problem is solved approximately.

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 refer the reader to Nilim and El Ghaoui [17] for the proof of the assertion that $\inf_{p \in \mathcal{P}(s,a)} \mathbf{E}^p[v]$ can be efficiently solved for this class of sets.

First consider the case where the underlying Markov chain is stationary. Suppose we have historical data consisting of triples $\mathcal{D} = \{(s_j, a_j, s'_j)\colon j \geq 1\}$, 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 $\hat{p}_{sa}$ of the conditional measure corresponding to the state-action pair (s, a) is given by

$$\hat{p}_{sa} = \operatorname*{arg\,max}_{p \in \mathcal{M}(\mathcal{S})} \left\{ \sum_{s' \in \mathcal{S}} n(s'|s, a) \log(p(s')) \right\}, \tag{33}$$

where

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

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

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

where $D(p_1 \| p_2)$ is the Kullback-Leibler or the relative entropy distance (see Chapter 2 in Cover and Thomas [7]) between two measures $p_1, p_2 \in \mathcal{M}(\mathcal{S})$ and is defined as follows:

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

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

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

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

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

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 $\hat{p}_{sa}$ is given by

$$\hat{p}_{sa} = \operatorname*{arg\,min}_{\{p \in \mathcal{M}(\mathcal{S}) \colon \mathbf{E}^p[g^j] = \bar{g}^j_{sa}, j = 1, \ldots, k\}} D(p \| p^0) \tag{37}$$

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

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 [7] for details) implies the following convergence in probability:

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

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 $\hat{p}_{sa}$ is, itself, a function of the sample size n_{sa}). Therefore,

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

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

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

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 [7]), $\mathcal{P}$ is convex for all $t \geq 0$.

Next, we handle the case of time-varying uncertainty sets $\mathcal{P}_t(s, a)$ introduced in §2. Since the uncertainty in this case is more refined, the data required to estimate the uncertainty structure needs to be more refined. We need data of the form $\mathcal{D} = \big\{(t_j, s_j, a_j, s'_j) \colon j \geq 1\big\}$ with the interpretation that state s'_j was observed in period $t_j + 1$ when the action a_j was employed in state s_j in period t_j. The first step is to "slice" the data according the period t to obtain $\mathcal{D}_t = \{(t_j, s_j, a_j, s'_j) \colon t_j = t\}$, $t \in T$. Next, the data $\mathcal{D}_t$ is used to estimate $\mathcal{P}_t(s, a)$ using the technique detailed above.

Relative entropy-based uncertainty sets were first introduced in Nilim and El Ghaoui [17]. We provide an information-theoretic rationale for employing these sets. The following result establishes that an ϵ-approximate solution for the robust problem corresponding to the set $\mathcal{P}$ in (39) can be computed efficiently.

LEMMA 4.1. *The value of the optimization problem*:

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

where $t > 0$, *is equal to*

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

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

We refer the reader to Nilim and El Ghaoui [17] for a proof of this lemma.

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

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

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

$$\mathcal{P} = \left\{p \in \mathcal{M}(\mathcal{S}) \colon \sum_{s\in\mathcal{S}}\frac{(p(s)-q(s))^2}{q(s)} \le t\right\}. \tag{42}$$

Sets of the form (42) were also introduced in Nilim and El Ghaoui [17]. Our new contribution is to show that this family of sets is a conservative approximation for the relative entropy-based sets. Next, we show that this approximation allows us to compute the *exact* value of the inner optimization problem.

LEMMA 4.2. *The value of the optimization problem*:

$$\begin{aligned} &\textit{minimize} \quad \mathbf{E}^p[v] \\ &\textit{subject to} \quad p \in \mathcal{P} = \left\{p \in \mathcal{M}(\mathcal{S}) \colon \sum_{s\in\mathcal{S}}\frac{(p(s)-q(s))^2}{q(s)} \le t\right\}, \end{aligned} \tag{43}$$

is equal to

$$\max_{\mu \ge \mathbf{0}}\left\{\mathbf{E}^q[v-\mu] - \sqrt{t\mathbf{Var}^q[v-\mu]}\right\}, \tag{44}$$

and the complexity of (44) *is* $\mathcal{O}(|\mathcal{S}|\log(|\mathcal{S}|))$.

PROOF. Let $y = p - q$. Then $p \in \mathcal{P}$ if and only if $\sum_s (y^2(s)/q(s)) \le t$, $\sum_s y(s) = 0$ and $y \ge -q$. Thus, the value of (43) is equal to

$$\begin{aligned} &\mathbf{E}^q[v] + \min \sum_s y(s)v(s), \\ &\text{subject to} \quad \sum_s \frac{y^2(s)}{q(s)} \le t, \\ &\qquad \sum_s y(s) = 0, \\ &\qquad y \ge -q. \end{aligned} \tag{45}$$

Lagrangian duality implies that the value of (45) is equal to

$$\begin{aligned}
&\mathbf{E}^q[v] + \max_{\mu\geq\mathbf{0},\,\gamma\geq\mathbf{0}} \min_{\{y:\,\sum_s y^2(s)/q(s)\leq t\}} \left\{ -\sum_{s\in\mathcal{S}} \mu(s)q(s) + \sum_{s\in\mathcal{S}} y(s)\big(v(s)-\gamma-\mu(s)\big) \right\} \\
&\quad = \max_{\mu\geq\mathbf{0},\,\gamma\in\mathbf{R}} \left\{ \mathbf{E}^q[v-\mu] - \sqrt{t\sum_{s\in\mathcal{S}} q(s)(v(s)-\mu(s)-\gamma)^2} \right\} \qquad (46) \\
&\quad = \max_{\mu\geq\mathbf{0}} \left\{ \mathbf{E}^q[v-\mu] - \sqrt{t\mathbf{Var}^q[v-\mu]} \right\}. \qquad (47)
\end{aligned}$$

This establishes that the value of (43) is equal to that of (44).

The optimum value of the inner minimization in (46) 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 $\mathcal{B}=\big\{s\in\mathcal{S}\colon \mu(s)>0\big\}$. Then complementary slackness conditions imply that $y^*(s)=-q(s)$, for all $s\in\mathcal{B}$, or equivalently,

$$v(s)-\mu(s) = \frac{\|\mathbf{z}\|}{\sqrt{t}} + \mathbf{E}^q[v-\mu] = \alpha, \quad \forall\, s\in\mathcal{B}; \tag{48}$$

i.e., $v(s)-\mu(s)$ is a constant for all $s\in\mathcal{B}$. Since the optimal value of (43) 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

$$\mu^*(s) = \begin{cases} v(s)-\alpha, & v(s)\geq\alpha, \\ 0, & \text{otherwise.} \end{cases} \tag{49}$$

Thus, dual optimization problem (44) reduces to solving for the optimal α. To this end, let $\{\hat{v}(k)\colon 1\leq k\leq|\mathcal{S}|\}$ denote the values $\{v(s)\colon s\in\mathcal{S}\}$ arranged in increasing order—an $\mathcal{O}(|\mathcal{S}|\log(|\mathcal{S}|))$ operation. Let $\hat{q}$ denote the sorted values of the measure q.

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

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

where $a_n=\sum_{k\leq n}\hat{q}(k)\hat{v}(k)$, $b_n=\sum_{k>n}\hat{q}(k)$, and $c_n=\sum_{k\leq n}\hat{q}(k)\hat{v}^2(k)$.

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

$$b_n^2\big(c_n+b_n\alpha^2-(a_n+b_n\alpha)^2\big) = t\big(b_n(1-b_n)\alpha-a_n\big)^2. \tag{50}$$

Thus, the optimal α can be computed by sequentially checking whether a root of (50) lies in $[\hat{v}_n,\hat{v}_{n+1})$, $n=1,\ldots,|\mathcal{S}|$. Let $a_0=c_0=0$ and $b_0=1$. Then

$$a_n=a_{n-1}+\hat{q}(n)\hat{v}(n), \quad b_n=b_{n-1}-\hat{q}_{n-1}, \quad c_n=c_{n-1}+\hat{q}(n)\hat{v}^2(n), \qquad n=1,\ldots,|\mathcal{S}|,$$

and computing the roots of (50) for a particular n is $\mathcal{O}(1)$ operation. Hence, computing the optimal α is $\mathcal{O}(|\mathcal{S}|)$ operation, and the overall complexity of computing a solution of (44) is $\mathcal{O}(|\mathcal{S}|\log(|\mathcal{S}|))$. □

Note that we compute the *exact* value (modulo finite precision errors) of the optimization problem (43). Nilim and El Ghaoui [17] show that the complexity of computing an ϵ-optimal solution for (43) is $\mathcal{O}(|\mathcal{S}|^{1.5}\log(v_{\max}/\epsilon))$.

From Lemma 12.6.1 in Cover and Thomas [7], 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

$$\mathcal{P} = \bigl\{p \colon \|p-q\|_1 \leq \sqrt{2\ln(2)t}\bigr\}, \tag{51}$$

is an outer approximation, i.e., relaxation, of the relative entropy uncertainty set (39). This approximation also allows us to compute the *exact* value of the inner optimization problem.

Lemma 4.3. *The value of the optimization problem*:

$$\begin{array}{ll} \textit{minimize} & \mathbf{E}^p[v] \\ \textit{subject to} & p \in \mathcal{P} = \bigl\{p\colon \|p-q\|_1 \leq \sqrt{2\ln(2)t}\bigr\}, \end{array} \tag{52}$$

is equal to

$$\mathbf{E}^q[v] - \frac{1}{2}\bigl(\sqrt{2\ln(2)t}\bigr)\min_{\mu\geq\mathbf{0}}\Bigl\{\Bigl(\max_s\{v(s)-\mu(s)\} - \min_s\{v(s)-\mu(s)\}\Bigr)\Bigr\}, \tag{53}$$

and the complexity of (53) *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 (52) is equal to

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

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

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

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 (44) reduces to solving for the optimal α. To this end, let $\{\hat{v}(k)\colon 1\leq k\leq|\mathcal{S}|\}$ denote the values $\{v(s)\colon s\in\mathcal{S}\}$ arranged in increasing order—an $\mathcal{O}(|\mathcal{S}|\log(|\mathcal{S}|))$ operation. Let $\hat{q}$ denote the sorted values of the measure q.

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

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

where $b_n = \sum_{k>n}\hat{q}_k$.

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

$$\alpha = \min\big\{\hat{v}(n)\colon\, b_n < \sqrt{2\ln(2)t}\big\}.$$

Since $b_1 = 1-\hat{q}_1$ and $b_n = b_{n-1} - \hat{q}_{n-1}$, $n \ge 2$, it follows that α can be computed in $\mathcal{O}(|\mathcal{S}|)$ time. □

In §4 we have, so far, been concerned with solving or suitably approximating the value of the inner optimization problem. The decision maker is, on the other hand, interested in approximating the value function $V(s)$. In the finite horizon case, the inner optimization problem is solved for each $n \in T$ and each state-action pair $(s_t, a_t) \in \mathcal{S}_t \times \mathcal{A}_t$. Hence, an approximation error in any stage n propagates all the way back to stage 0. In the infinite horizon setting, the approximation error propagates via the value iteration step. Consequently, in order to compute an ϵ-optimal solution for the value function, one has to carefully set the level of approximation for the inner optimization problem. We refer readers interested in this problem to Nilim and El Ghaoui [17] where the authors provide a careful analysis of the error propagation.

5. Games with perfect information. In this section we show that robust DP with dynamic uncertainty is equivalent to zero-sum games with perfect information (Gillette [13], Altman et al. [1]). We will restrict attention to finite horizon robust DP. We assume that there are $N-1$ decision epochs with the reward at epoch N given by the function r_N. For simplicity of exposition, we assume that the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are not functions of time. The extension to discounted infinite horizon games is straightforward.

Consider a two-person zero-sum stochastic game with a countable state space $\tilde{\mathcal{S}}$. Let $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ denote the metric spaces of actions for players 1 and 2 respectively. Let $p(\cdot \mid s, a, b) \in \mathcal{M}(\tilde{\mathcal{S}})$ denote the transition probability when in state $s \in \tilde{\mathcal{S}}$ player 1 chooses action $a \in \tilde{\mathcal{A}}(s)$ and player 2 chooses action $b \in \tilde{\mathcal{B}}(s)$. Let $r_t\colon \tilde{\mathcal{S}} \times \tilde{\mathcal{A}} \times \tilde{\mathcal{B}}$ denote the reward function at time t. A stochastic game is said to have *perfect information* if the state space $\tilde{\mathcal{S}}$ can be partitioned as $\tilde{\mathcal{S}} = \tilde{\mathcal{S}}_1 \cup \tilde{\mathcal{S}}_2$, $\tilde{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_2 = \varnothing$, such that the action set for player 1, $\tilde{\mathcal{A}}(s)$, is a singleton for all $s \in \tilde{\mathcal{S}}_2$ and the action set of player 2, $\tilde{\mathcal{B}}(s)$, is a singleton for all $s \in \tilde{\mathcal{S}}_1$.

A special case of games with perfect information are games where player 2 chooses the action $b \in \tilde{\mathcal{B}}(s)$ after observing the action $a \in \tilde{\mathcal{A}}(s)$ of player 1. Robust dynamic programming with dynamic model for uncertainty is a game of this form: Player 1 is the decision maker who chooses an action $a \in \mathcal{A}(s)$ and player 2 is the adversary who chooses the state transition $p_{sa} \in \mathcal{P}(s,a)$ after observing $a \in \mathcal{A}(s)$. Robust dynamic programming can be reduced to a game with perfect information as follows:

(i) Define $\tilde{\mathcal{S}} = \mathcal{S} \cup (\mathcal{S} \times \mathcal{A})$.

(ii) At the decision epoch $t = 0$, the state $s \in \mathcal{S} \subset \tilde{\mathcal{S}}$.

(iii) For all $s \in \mathcal{S} \subset \tilde{\mathcal{S}}$, define

$$\begin{aligned} \tilde{\mathcal{A}}(s) &= \mathcal{A}(s), \\ \tilde{\mathcal{B}}(s) &= \{\rho\}, \end{aligned} \tag{54}$$

where ρ denotes an arbitrary singleton.

(iv) For all $(s,a)\in\mathcal{S}\times\mathcal{A}\subset\widetilde{\mathcal{S}}$, define

$$\begin{aligned}\tilde{\mathcal{A}}(s,a)&=\{\rho\}, \quad \text{for some } \rho,\\ \widetilde{\mathcal{B}}(s,a)&=\mathcal{P}(s,a),\end{aligned} \tag{55}$$

where ρ denotes an arbitrary singleton.

(v) Define the state transition $p_t(\tilde{s}'|\tilde{s},\tilde{a},\tilde{b})$ as follows.

$$p_t(\tilde{s}'\mid\tilde{s},\tilde{a},\tilde{b})=\begin{cases}1 & \tilde{s}=s\in\mathcal{S},\, \tilde{a}\in\mathcal{A}(s),\, \tilde{b}=\rho,\, \tilde{s}'=(s,\tilde{a})\in\mathcal{S}\times\mathcal{A},\\ & \quad t=2n,\ 0\le n\le N-1,\\ p_{sa} & \tilde{s}=(s,a)\in\mathcal{S}\times\mathcal{A},\, \tilde{a}=\rho,\, \tilde{b}=p_{sa}\in\mathcal{P}(s,a),\\ & \quad t=2n+1,\ 0\le n\le N-1,\\ 0 & \text{otherwise.}\end{cases} \tag{56}$$

(vi) The reward function $r_t(\tilde{s},\tilde{a},\tilde{b})$ is defined as follows.

$$r_t(\tilde{s},\tilde{a},\tilde{b})=\begin{cases}r_t(s,\tilde{a}) & \tilde{s}=s\in\mathcal{S},\, \tilde{a}\in\mathcal{A}(s),\, \tilde{b}=\rho,\, t=2n,\ 0\le n\le N-1,\\ r_N(s), & t=2N,\\ 0 & \text{otherwise.}\end{cases} \tag{57}$$

We assume here that the reward function of the robust dynamic program only depends on the current state-action pair. The extension to reward functions of the form $r(s,a,s')$ is straightforward.

The game constructed by (i)–(vi) above proceeds as follows. In state $s\in\mathcal{S}$, the decision maker chooses an action $a\in\mathcal{A}(s)$ and receives a reward $r(s,a)$. The new state of the system is now $(s,a)\in\mathcal{S}\times\mathcal{A}$. In this state, the adversary chooses $p_{sa}\in\mathcal{P}(s,a)$ that determines the new state $s'\in\mathcal{S}$. This state transition does not yield any reward to the decision maker. Thus, the game constructed above is a sequential game and each decision-making stage in the robust DP corresponds to two sequential stages in the stochastic game. We will denote the stages corresponding to the nth robust DP stage by the pair $(2n,2n+1)$. In order to define the value of this stochastic two-person game we need to make the space of actions convex by allowing randomized actions. In state $s\in\mathcal{S}\subset\widetilde{\mathcal{S}}$ a randomized action for Player 1 is given, as before, by a probability measure $q\in\mathcal{M}(\mathcal{A}(s))$. In state $(s,a)\in\mathcal{S}\times\mathcal{A}\subset\widetilde{\mathcal{S}}$ a randomized action for Player 2, i.e., the adversary, is given by a measure $p\in\mathcal{M}(\mathcal{P}(s,a))=\mathbf{conv}(\mathcal{P}(s,a))$, the convex hull of the set $\mathcal{P}(s,a)$. Note that we have implicitly made the set of transition measures $\mathcal{P}(s,a)$ convex.

Standard results in the theory of zero-sum games (see, e.g., Nowak [19]) imply that the value functions $\{\widetilde{V}_k\colon k=0,\ldots,2N\}$ of the finite horizon zero-sum stochastic game constructed above is characterized by the recursion

$$\begin{aligned}\widetilde{V}_k(\tilde{s})&=\max_{q\in\mathcal{M}(\tilde{A}(\tilde{s}))}\left\{\min_{p\in\mathcal{M}(\widetilde{\mathcal{B}}(\tilde{s}))}[\mathbf{E}^{qp}[r(\tilde{s},a,b)+\widetilde{V}_{k+1}(\tilde{s}')]]\right\}, \quad 0\le k\le 2N-1,\, \tilde{s}\in\widetilde{\mathcal{S}},\\ \widetilde{V}_{2N}(s)&=r_N(s), \quad s\in\mathcal{S}\subset\widetilde{\mathcal{S}},\end{aligned} \tag{58}$$

where

$$\mathbf{E}^{qp}[r(s,a,b)+\widetilde{V}_{k+1}(\tilde{s}')]=\int_{\mathcal{A}(s)}\int_{\mathcal{B}(s)}\left(r(s,a,b)+\int_{\widetilde{\mathcal{S}}}\widetilde{V}_{k+1}(\tilde{s}')\,p(d\tilde{s}'\mid\tilde{s},a,b)\right)q(da)\,p(db).$$

From the dynamics described in (54)–(57) and the fact that the game is sequential, it follows that recursion (58) can be simplified as follows:

$$\begin{aligned}\widetilde{V}_{2n}(s)&=\max_{q\in\mathcal{M}(\tilde{A}(s))}\mathbf{E}^q\Big[r(\tilde{s},a,b)+\min_{p\in\mathbf{conv}(\mathcal{P}(s,a))}\mathbf{E}^p[\widetilde{V}_{2(n+1)}]\Big], \quad n=1,\ldots,N-1,\, \tilde{s}\in\mathcal{S},\\ \widetilde{V}_{2N}(s)&=r_N(s), \quad s\in\mathcal{S}.\end{aligned} \tag{59}$$

From the construction of the stochastic game, we have that the value function $V_n(s)$ of the robust DP can be identified with $\widetilde{V}_{2n}(s)$, i.e., $V_n(s) = \widetilde{V}_{2n}(s)$. Thus, we have that

$$V_n(s) = \max_{q \in \mathcal{M}(\tilde{A}(s))} \mathbf{E}^q \Big[r(\tilde{s}, a, b) + \min_{p \in \mathbf{conv}(\mathcal{P}(s,a))} \mathbf{E}^p [V_{n+1}] \Big].$$

Since $\mathbf{E}^p[V_{n+1}]$ is a linear function of the measure p, it follows that

$$\min_{p \in \mathbf{conv}(\mathcal{P}(s,a))} \mathbf{E}^p [V_{n+1}] = \min_{p \in \mathcal{P}(s,a)} \mathbf{E}^p [V_{n+1}].$$

An argument similar to the one used to establish (15) in Theorem 2.1 establishes that the decision maker can be restricted to deterministic actions. Thus, we recover the robust Bellman recursion

$$V_n(s) = \max_{a \in \tilde{A}(s)} \min_{p \in \mathcal{P}(s,a)} \mathbf{E}^p [r(\tilde{s}, a, b) + V_{n+1}].$$

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. Since these confidence regions are described in terms of the relative entropy or the Kullback-Liebler distance, we are naturally led to the uncertainty sets discussed in §4. This family of sets was first introduced in Nilim and El Ghaoui [17].

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 the 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 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 nonrobust DP. Although independently obtained, these results were first obtained in Nilim and El Ghaoui [17]. 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.

The robust value function V^* provides a lower bound on the achievable performance; one can also define an optimistic value function $\overline{V}^*$ 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.

As indicated in the introduction, we restricted our attention to problems where the nonrobust 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 [5]).

Appendix A. Consequences of rectangularity.

We will begin with an example that illustrates the inappropriateness of rectangularity in a finite horizon setting. This example is a dynamic version of the Ellsberg Urn problem (Ellsberg [9]) discussed in Epstein and Schneider [10].

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 color of the ball is revealed at time $t = 2$. At the

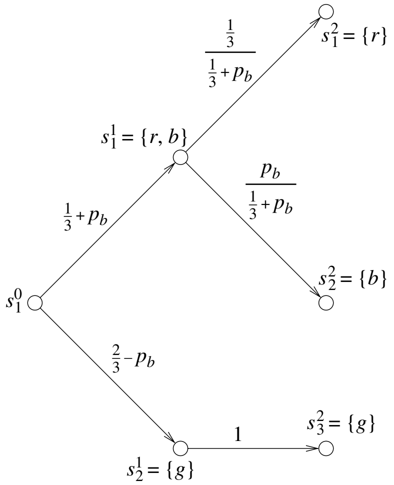

FIGURE 3. Dynamic Ellsberg experiment.

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 3 where $p_b = \mathbf{P}\{\text{ball is blue}\}$.

Suppose $p_b \in [\underline{p}, \bar{p}] \subseteq [0, 2/3]$ is ambiguous. Consider the robust optimal stopping problem where the state transition is given by Figure 3. In each state $s \in \mathcal{S}_t$ at time $t = 0, 1$, there 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 3 implies that the conditional measures consistent with the decision rules $\bar{d}_i$, $i = 0, 1$, are given by

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

Thus,

$$\mathcal{T}^{\bar{d}_0} \times \mathcal{T}^{\bar{d}_1} = \left\{\begin{matrix}\left(p(s_1^1 \mid s_1^0), p(s_2^1 \mid s_1^0)\right) = (1/3+\alpha, 2/3-\alpha), & \\ \left(p(s_1^2 \mid s_1^1), p(s_2^2 \mid s_1^1)\right) = \left(\dfrac{1/3}{1/3+\alpha'}, \dfrac{\alpha'}{1/3+\alpha'}\right), p(s_2^2 \mid s_2^1) = 1 & \colon \alpha, \alpha' \in [\underline{p}, \bar{p}]\end{matrix}\right\},$$

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

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

The problem arises because the information structure in Figure 3 assumes that there is a single urn that decides the 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. In spite 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 an infinite horizon setting where a state-action can be revisited. In this setting rectangularity may not be appropriate when 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 3.3 shows that the problems with rectangularity disappear if one restricts the decision maker to stationary policies.

Acknowledgments. The author thanks the anonymous referees for insightful comments that have substantially improved the presentation and the conceptual content of the paper. The author is grateful to Prof. Eugene A. Feinberg for bringing the connection between robust DP and stochastic games with perfect information to his attention. The author's research was partially supported by NSF Grants CCR-00-09972 and DMS-01-04282, as well as by ONR Grant N000140310514.

References

[1] Altman, E., E. A. Feinberg, A. Shwartz. 2000. Weighted discounted games with perfect information. J. A. Filar, V. Gaitsgory, K. Mizukami, eds. *Advances in Dynamic Games and Applications*. Birkhauser, Boston, MA, 303–323.

[2] Bagnell, J., A. Ng, J. Schneider. 2001. Solving uncertain Markov decision problems. Technical Report, Robotics Institute, CMU, CMU-RI-TR-01-25.

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

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

[5] Bertsekas, D. P., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

[6] Cover, T. M. 1991. Universal portfolios. *Math. Finance* **1**(1) 1–29.

[7] Cover, T. M., J. A. Thomas. 1991. *Elements of Information Theory*. John Wiley and Sons, New York.

[8] de Farias, D. P., B. Van Roy. 2003. The linear programming approach to approximate dynamic programming. *Oper. Res.* **51** 850–865.

[9] Ellsberg, D. 1961. Risk, ambiguity and the Savage axioms. *Quart. J. Econom.* **25**(4) 643–669.

[10] Epstein, L. G., M. Schneider. 2001. Recursive multiple priors. *J. Econom. Theory* **113**(1) 1–31. http://rcer.econ.rochester.edu.

[11] Epstein, L. G., M. Schneider. 2002. Learning under ambiguity. Technical Report 497, Rochester Center for Economic Research. http://rcer.econ.rochester.edu.

[12] Gilboa, I., D. Schmeidler. 1989. Maxmin expected utility with non-unique priors. *J. Math. Econom.* **18** 141–153.

[13] Gillette, D. 1957. Stochastic games with zero stop probabilities. M. Dresher, A. W. Tucker, P. Wolfe, eds. *Contributions to the Theory of Games, Vol. 3*. Princeton University Press, Princeton, NJ, 179–187.

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

[15] Hansen, L. P., T. J. Sargent. 2001. Robust control and model uncertainty. *Amer. Econom. Rev.* **91** 60–66.

[16] Littman, M. 1994. Memoryless policies: Theoretical limitations and practical results. D. Cliff, P. Husbands, S. W. Wilson, eds. *Animals to Animats: SAB '94*. MIT Press, Cambridge, MA, 238–245.

[17] Nilim, A., L. El Ghaoui. 2002. Robust solutions to Markov decision problems with uncertain transition matrices. *Oper. Res.* Forthcoming.

[18] Nilim, A., L. El Ghaoui. 2003. Robustness in Markov decision problems with uncertain transition matrices. *Proc. NIPS*.

[19] Nowak, A. S. 1984. On zero-sum stochastic games with general state space. I. *Probab. Math. Statist.* **4** 13–32.

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

[21] Satia, J. K. 1968. Markovian decision process with uncertain transition matrices or/and probabilistic observation of states. Ph.D. thesis, Stanford University, Stanford, CA.

[22] Satia, J. K., R. L. Lave. 1973. Markov decision processes with uncertain transition probabilities. *Oper. Res.* **21**(3) 728–740.

[23] Shapiro, A., A. J. Kleywegt. 2002. Minimax analysis of stochastic problems. *Optim. Methods and Software* Forthcoming.

[24] Simester, D. I., P. Sun, J. Tsitsiklis. 2005. Dynamic catalog mailing policies. Technical Report FRPS05-202, Fuqua School of Business, Duke University, Durham, NC.

[25] White, C. C., H. K. Eldieb. 1994. Markov decision processes with imprecise transition probabilities. *Oper. Res.* **43** 739–749.