Robust Solutions to Markov Decision Problems

Arnab Nilim and Laurent El Ghaoui
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA 94720
nilim@eecs.berkeley.edu, elghaoui@eecs.berkeley.edu

Abstract

Optimal solutions to Markov Decision Problems (MDPs) are very sensitive with respect to the state transition probabilities. In many practical problems, the estimation of those probabilities is far from accurate. Hence, estimation errors are limiting factors in applying MDPs to real-world problems. We propose an algorithm for solving £nite-state and £nite-action MDPs, where the solution is guaranteed to be robust with respect to estimation errors on the state transition probabilities. Our algorithm involves a statistically accurate yet numerically ef£cient representation of uncertainty via likelihood functions. The worst-case complexity of the robust algorithm is the same as the original Bellman recursion. Hence, robustness can be added at practically no extra computing cost.

Introduction

We consider a £nite-state and £nite-action Markov decision problem in which the transition probabilities themselves are uncertain, and seek a robust decision for it. Our work is motivated by the fact that in many practical problems, the transition matrices have to be estimated from data. This may be a dif£cult task and the estimation errors may have a huge impact on the solution, which is often quite sensitive to changes in the transition probabilities (Feinberg & Shwartz 2002). A number of authors have addressed the issue of uncertainty in the transition matrices of an MDP. A Bayesian approach such as described by (Shapiro & Kleywegt 2002) requires a perfect knowledge of the whole prior distribution on the transition matrix, making it dif£cult to apply in practice. Other authors have considered the transition matrix to lie in a given set, most typically a polytope: see (Satia & Lave 1973; White & Eldeib 1994; Givan, Leach, & Dean 1997). Although our approach allows to describe the uncertainty on the transition matrix by a polytope, we may argue *against* choosing such a model for the uncertainty. First, a general polytope is often not a tractable way to address the robustness problem, as it incurs a signi£cant additional computational effort to handle uncertainty. Perhaps more importantly, polytopic models, especially interval matrices, may be very poor representations of statistical uncertainty and lead to very conservative robust policies. In (Bagnell, Ng, & Schneider 2001), the authors consider a problem dual to ours, and provide a general statement according to which the cost of solving their problem is polynomial in problem size, provided the uncertainty on the transition matrices is described by convex sets, without proposing any speci£c algorithm. This paper is a short version of a longer report (Nilim & El-Ghaoui 2004), which contains all the proofs of the results summarized here.

Notation. $P > 0$ or $P \geq 0$ refers to the strict or non-strict componentwise inequality for matrices or vectors. For a vector $p > 0$, $\log p$ refers to the componentwise operation. The notation $\mathbf{1}$ refers to the vector of ones, with size determined from context. The probability simplex in $\mathbf{R}^n$ is denoted $\Delta_n = \{p \in \mathbf{R}^n_+ \,:\, p^T\mathbf{1} = 1\}$, while Θ_n is the set of $n \times n$ transition matrices (componentwise non-negative matrices with rows summing to one). We use $\sigma_{\mathcal{P}}$ to denote the support function of a set $\mathcal{P} \subseteq \mathbf{R}^n$, with for $v \in \mathbf{R}^n$, $\sigma_{\mathcal{P}}(v) := \sup\{p^T v \,:\, p \in \mathcal{P}\}$.

The problem description

We consider a £nite horizon Markov decision process with £nite decision horizon $T = \{0, 1, 2, \ldots, N-1\}$. At each stage, the system occupies a state $i \in \mathcal{X}$, where $n = |\mathcal{X}|$ is £nite, and a decision maker is allowed to choose an action a deterministically from a £nite set of allowable actions $\mathcal{A} = \{a_1, \ldots, a_m\}$ (for notational simplicity we assume that $\mathcal{A}$ is not state-dependent). The system starts in a given initial state i_0. The states make Markov transitions according to a collection of (possibly time-dependent) transition matrices $\tau := (P_t^a)_{a \in \mathcal{A}, t \in T}$, where for every $a \in \mathcal{A}$, $t \in T$, the $n \times n$ transition matrix P_t^a contains the probabilities of transition under action a at stage t. We denote by $\pi = (\mathbf{a}_0, \ldots, \mathbf{a}_{N-1})$ a generic controller policy, where $\mathbf{a}_t(i)$ denotes the controller action when the system is in state $i \in \mathcal{X}$ at time $t \in T$. Let $\Pi = \mathcal{A}^{nN}$ be the corresponding strategy space. De£ne by $c_t(i, a)$ the cost corresponding to state $i \in \mathcal{X}$ and action $a \in \mathcal{A}$ at time $t \in T$, and by c_N the cost function at the terminal stage. We assume that $c_t(i, a)$ is non-negative and £nite for every $i \in \mathcal{X}$ and $a \in \mathcal{A}$.

For a given set of transition matrices τ, we de£ne the

£nite-horizon *nominal* problem by

$$\phi_N(\Pi, \tau) := \min_{\pi \in \Pi} C_N(\pi, \tau), \tag{1}$$

where $C_N(\pi, \tau)$ denotes the *expected total cost* under controller policy π and transitions τ:

$$C_N(\pi, \tau) := \mathbf{E}\left(\sum_{t=0}^{N-1} c_t(i_t, \mathbf{a}_t(i)) + c_N(i_N)\right). \tag{2}$$

A special case of interest is when the expected total cost function bears the form (2), where the terminal cost is zero, and $c_t(i, a) = \nu^t c(i, a)$, with $c(i, a)$ now a constant cost function, which we assume non-negative and £nite everywhere, and $\nu \in (0, 1)$ is a discount factor. We refer to this cost function as the discounted cost function, and denote by $C_\infty(\pi, \tau)$ the limit of the discounted cost (2) as $N \to \infty$.

When the transition matrices are exactly known, the corresponding nominal problem can be solved via a dynamic programming algorithm, which has total complexity of mn^2N ¤ops in the £nite-horizon case. In the in£nite-horizon case with a discounted cost function, the cost of computing an ϵ-suboptimal policy via the Bellman recursion is $O(mn^2 \log(1/\epsilon))$; see (Puterman 1994) for more details.

The robust control problems

At £rst we assume that when for each action a and time t, the corresponding transition matrix P_t^a is only known to lie in some given subset $\mathcal{P}^a$. Two models for transition matrix uncertainty are possible, leading to two possible forms of £nite-horizon robust control problems. In a £rst model, referred to as the *stationary uncertainty* model, the transition matrices are chosen by nature depending on the controller policy once and for all, and remain £xed thereafter. In a second model, which we refer to as the *time-varying uncertainty* model, the transition matrices can vary arbitrarily with time, within their prescribed bounds. Each problem leads to a game between the controller and nature, where the controller seeks to minimize the maximum expected cost, with nature being the maximizing player.

Let us de£ne our two problems more formally. A *policy of nature* refers to a speci£c collection of time-dependent transition matrices $\tau = (P_t^a)_{a \in \mathcal{A}, t \in T}$ chosen by nature, and the set of admissible policies of nature is $\mathcal{T} := (\otimes_{a \in \mathcal{A}} \mathcal{P}^a)^N$. Denote by $\mathcal{T}_s$ the set of stationary admissible policies of nature:
$\mathcal{T}_s = \{\tau = (P_t^a)_{a \in \mathcal{A}, t \in T} \in \mathcal{T} \ : \ P_t^a = P_s^a$ for every $t, s \in T$, $a \in \mathcal{A}\}$.
The stationary uncertainty model leads to the problem

$$\phi_N(\Pi, \mathcal{T}_s) := \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}_s} C_N(\pi, \tau). \tag{3}$$

In contrast, the time-varying uncertainty model leads to a relaxed version of the above:

$$\phi_N(\Pi, \mathcal{T}_s) \leq \phi_N(\Pi, \mathcal{T}) := \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} C_N(\pi, \tau). \tag{4}$$

The £rst model is attractive for statistical reasons, as it is much easier to develop statistically accurate sets of con£dence when the underlying process is time-invariant. Unfortunately, the resulting game (3) seems to be hard to solve because "principle of optimality" may not hold in these problems due to the dependence of optimal actions of nature among different stages. The second model is attractive as one can solve the corresponding game (4) using a variant of the dynamic programming algorithm seen later, but we are left with a dif£cult task, that of estimating a meaningful set of con£dence for the time-varying matrices P_t^a. In this paper we will use the £rst model of uncertainty in order to derive statistically meaningful sets of con£dence for the transition matrices, based on likelihood or entropy bounds. Then, instead of solving the corresponding dif£cult control problem (3), we use an approximation that is common in robust control, and solve the time-varying upper bound (4), using the uncertainty sets $\mathcal{P}^a$ derived from a stationarity assumption about the transition matrices. We will also consider a variant of the £nite-horizon time-varying problem (4), where controller and nature play alternatively, leading to a repeated game

$$\phi_N^{\text{rep}}(\Pi, \mathcal{Q}) := \min_{\mathbf{a}_0} \max_{\tau_0 \in \mathcal{Q}} \min_{\mathbf{a}_1} \max_{\tau_1 \in \mathcal{Q}} \ldots \min_{\mathbf{a}_{N-1}} \max_{\tau_{N-1} \in \mathcal{Q}} C_N(\pi, \tau), \tag{5}$$

where the notation $\tau_t = (P_t^a)_{a \in \mathcal{A}}$ denotes the collection of transition matrices at a given time $t \in T$, and $\mathcal{Q} := \otimes_{a \in \mathcal{A}} \mathcal{P}^a$ is the corresponding set of con£dence.

Finally, we will consider an in£nite-horizon robust control problem, with the discounted cost function referred to above, and where we restrict control and nature policies to be stationary:

$$\phi_\infty(\Pi_s, \mathcal{T}_s) := \min_{\pi \in \Pi_s} \max_{\tau \in \mathcal{T}_s} C_\infty(\pi, \tau), \tag{6}$$

where Π_s denotes the space of stationary control policies. We de£ne $\phi_\infty(\Pi, \mathcal{T})$, $\phi_\infty(\Pi, \mathcal{T}_s)$ and $\phi_\infty(\Pi_s, \mathcal{T})$ accordingly.

In the sequel, for a given control policy $\pi \in \Pi$ and subset $\mathcal{S} \subseteq \mathcal{T}$, the notation $\phi_N(\pi, \mathcal{S}) := \max_{\tau \in \mathcal{S}} C_N(\pi, \tau)$ denotes the worst-case expected total cost for the £nite-horizon problem, and $\phi_\infty(\pi, \mathcal{S})$ is de£ned likewise.

Main results

Our main contributions are as follows. First we provide a recursion, the "robust dynamic programming" algorithm, which solves the £nite-horizon robust control problem (4). We provide a simple proof in (Nilim & El-Ghaoui 2004) of the optimality of the recursion, where the main ingredient is to show that perfect duality holds in the game (4). As a corollary of this result, we obtain that the repeated game (5) is equivalent to its non-repeated counterpart (4). Second, we provide similar results for the in£nite-horizon problem with discounted cost function, (6). Moreover, we obtain that if we consider a £nite-horizon problem with a discounted cost function, then the gap between the optimal value of the stationary uncertainty problem (3) and that of its time-varying counterpart (4) goes to zero as the horizon length goes to in£nity, at a rate determined by the discount factor. Finally, we identify several classes of uncertainty models, which result in an algorithm that is *both* statistically accurate and numerically tractable. We provide precise complexity results that imply that, with the proposed approach, robustness can be handled at practically no extra computing cost.

Finite-Horizon robust MDP

We consider the £nite-horizon robust control problem de£ned in section . For a given state $i \in \mathcal{X}$, action $a \in \mathcal{A}$, and $P^a \in \mathcal{P}^a$, we denote by p_i^a the next-state distribution drawn from P^a corresponding to state $i \in \mathcal{X}$; thus p_i^a is the i-th row of matrix P^a. We de£ne $\mathcal{P}_i^a$ as the projection of the set $\mathcal{P}^a$ onto the set of p_i^a-variables. By assumption, these sets are included in the probability simplex of $\mathbf{R}^n$, Δ_n; no other property is assumed. The following theorem is proved in (Nilim & El-Ghaoui 2004).

Theorem 1 (robust dynamic programming) *For the robust control problem (4), perfect duality holds:*

$$\phi_N(\Pi, \mathcal{T}) = \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} C_N(\pi, \tau) = \max_{\tau \in \mathcal{T}} \min_{\pi \in \Pi} C_N(\pi, \tau)$$

$$:= \psi_N(\Pi, \mathcal{T}).$$

The problem can be solved via the recursion

$$v_t(i) = \min_{a \in \mathcal{A}} \left(c_t(i,a) + \sigma_{\mathcal{P}_i^a}(v_{t+1}) \right), \quad i \in \mathcal{X}, \quad t \in T, \quad (7)$$

where $\sigma_{\mathcal{P}}(v) := \sup\{p^T v \ : \ p \in \mathcal{P}\}$ denotes the support function of a set $\mathcal{P}$, $v_t(i)$ is the worst-case optimal value function in state i at stage t. A corresponding optimal control policy $\pi^* = (\mathbf{a}_0^*, \ldots, \mathbf{a}_{N-1}^*)$ is obtained by setting

$$\mathbf{a}_t^*(i) \in \arg\min_{a \in \mathcal{A}} \left\{ c_t(i,a) + \sigma_{\mathcal{P}_i^a}(v_{t+1}) \right\}, \quad i \in \mathcal{X}. \quad (8)$$

The effect of uncertainty on a given *strategy $\pi = (\mathbf{a}_0, \ldots, \mathbf{a}_N)$ can be evaluated by the following recursion*

$$v_t^\pi(i) \quad = \quad c_t(i, \mathbf{a}_t(i)) + \sigma_{\mathcal{P}_i^{\mathbf{a}_t(i)}}(v_{t+1}^\pi), \quad i \in \mathcal{X}, \quad (9)$$

which provides the worst-case value function v^π for the strategy π.

The above result has a nice consequence for the repeated game (5):

Corollary 2 *The repeated game (5) is equivalent to the game (4):*

$$\phi_N^{\text{rep}}(\Pi, \mathcal{Q}) = \phi_N(\Pi, \mathcal{T}),$$

and the optimal strategies for $\phi_N(\Pi, \mathcal{T})$ given in theorem 1 are optimal for $\phi_N^{\text{rep}}(\Pi, \mathcal{Q})$ as well.

The interpretation of the perfect duality result given in theorem 1, and its consequence given in corollary 2, is that it does not matter wether the controller or nature play £rst, or if they alternatively; all these games are equivalent. Now consider the following algorithm, where the uncertainty is described in terms of one of the models described in section "Kullback-Liebler Divergence Uncertainty Models":

Robust Finite Horizon Dynamic Programming Algorithm

1. *Set $\epsilon > 0$. Initialize the value function to its terminal value $\hat{v}_N = c_N$.*
2. *Repeat until $t = 0$:*
   - *(a) For every state $i \in \mathcal{X}$ and action $a \in \mathcal{A}$, compute, using the bisection algorithm given in (Nilim & El-Ghaoui 2004), a value $\hat{\sigma}_i^a$ such that*
   $$\hat{\sigma}_i^a - \epsilon/N \leq \sigma_{\mathcal{P}_i^a}(\hat{v}_t) \leq \hat{\sigma}_i^a.$$
   - *(b) Update the value function by $\hat{v}_{t-1}(i) = \min_{a \in \mathcal{A}}(c_{t-1}(i,a) + \hat{\sigma}_i^a)$, $i \in \mathcal{X}$.*
   - *(c) Replace t by $t-1$ and go to 2.*
3. *For every $i \in \mathcal{X}$ and $t \in T$, set $\pi^\epsilon = (\mathbf{a}_0^\epsilon, \ldots, \mathbf{a}_{N-1}^\epsilon)$, where*
   $$\mathbf{a}_t^\epsilon(i) = \arg\max_{a \in \mathcal{A}} \left\{ c_{t-1}(i,a) + \hat{\sigma}_i^a \right\}, \quad i \in \mathcal{X}, \quad a \in \mathcal{A}.$$

As shown in (Nilim & El-Ghaoui 2004), the above algorithm provides an suboptimal policy π^ϵ that achieves the exact optimum with prescribed accuracy ϵ, with a required number of ¤ops bounded above by $O(mn^2 N \log(N/\epsilon))$. This means that robustness is obtained at a relative increase of computational cost of only $\log(N/\epsilon)$ with respect to the classical dynamic programming algorithm, which is small for moderate values of N. If N is very large, we can turn instead to the in£nite-horizon problem examined in the following section, and similar complexity results hold.

In£nite-Horizon MDP

In this section, we address the in£nite-horizon robust control problem, with a discounted cost function of the form (2), where the terminal cost is zero, and $c_t(i,a) = \nu^t c(i,a)$, where $c(i,a)$ is now a constant cost function, which we assume non-negative and £nite everywhere, and $\nu \in (0,1)$ is a discount factor.

We begin with the in£nite-horizon problem involving stationary control and nature policies de£ned in (6). The following theorem is proved in (Nilim & El-Ghaoui 2004).

Theorem 3 (Robust Bellman recursion) *For the in£nite-horizon robust control problem (6) with stationary uncertainty on the transition matrices, stationary control policies, and a discounted cost function with discount factor $\nu \in [0,1)$, perfect duality holds:*

$$\phi_\infty(\Pi_s, \mathcal{T}_s) = \max_{\tau \in \mathcal{T}_s} \min_{\pi \in \Pi_s} C_\infty(\pi, \tau) := \psi_\infty(\Pi_s, \mathcal{T}_s). \quad (10)$$

The optimal value is given by $\phi_\infty(\Pi_s, \mathcal{T}_s) = v(i_0)$, where i_0 is the initial state, and where the value function v satis£es the optimality conditions

$$v(i) = \min_{a \in \mathcal{A}} \left(c(i,a) + \nu \sigma_{\mathcal{P}_i^a}(v) \right), \quad i \in \mathcal{X}. \quad (11)$$

The value function is the unique limit value of the convergent vector sequence de£ned by

$$v_{k+1}(i) = \min_{a \in \mathcal{A}} \left(c(i,a) + \nu \sigma_{\mathcal{P}_i^a}(v_k) \right), \quad i \in \mathcal{X}, \quad k = 1, 2, \ldots \quad (12)$$

A stationary, optimal control policy $\pi = (\mathbf{a}^*, \mathbf{a}^*, \ldots)$ is obtained as

$$\mathbf{a}^*(i) \in \arg\min_{a \in \mathcal{A}} \left\{ c(i,a) + \nu \sigma_{\mathcal{P}_i^a}(v) \right\}, \quad i \in \mathcal{X}. \quad (13)$$

Note that the problem of computing the dual quantity $\psi_\infty(\Pi_s, \mathcal{T}_s)$ given in (10), has been addressed in (Bagnell, Ng, & Schneider 2001), where the authors provide the recursion (12) without proof.

Theorem (3) leads to the following corollary, also proved in (Nilim & El-Ghaoui 2004).

Corollary 4 *In the in£nite-horizon problem, we can without loss of generality assume that the control and nature policies are stationary, that is,*

$$\phi_\infty(\Pi, \mathcal{T}) = \phi_\infty(\Pi_s, \mathcal{T}_s) = \phi_\infty(\Pi_s, \mathcal{T}) = \phi_\infty(\Pi, \mathcal{T}_s). \tag{14}$$

Furthermore, in the £nite-horizon case, with a discounted cost function, the gap between the optimal values of the £nite-horizon problems under stationary and time-varying uncertainty models, $\phi_N(\Pi, \mathcal{T}) - \phi_N(\Pi, \mathcal{T}_s)$, goes to zero as the horizon length N goes to in£nity, at a geometric rate ν.

Now consider the following algorithm, where we describe the uncertainty using one of the models of section "Kullback-Liebler Divergence Uncertainty Models".

Robust In£nite Horizon Dynamic Programming Algorithm

1. *Set $\epsilon > 0$, initialize the value function $\hat{v}_1 > 0$ and set $k = 1$.*

2(a) *For all states i and controls a, compute, using the bisection algorithm given in (Nilim & El-Ghaoui 2004), a value $\hat{\sigma}_i^a$ such that*

$$\hat{\sigma}_i^a - \delta \leq \sigma_{\mathcal{P}_i^a}(\hat{v}_k) \leq \hat{\sigma}_i^a,$$

where $\delta = (1-\nu)\epsilon/2\nu$.

(b) *For all states i and controls a, compute $\hat{v}_{k+1}(i)$ by,*

$$\hat{v}_{k+1}(i) = \min_{a \in \mathcal{A}} \left(c(i,a) + \nu \hat{\sigma}_i^a \right).$$

3. *If*

$$\|\hat{v}_{k+1} - \hat{v}_k\| < \frac{(1-\nu)\epsilon}{2\nu},$$

go to 4. Otherwise, replace k by $k+1$ and go to 2.

4. *For each $i \in \mathcal{X}$, set an $\pi^\epsilon = (\mathbf{a}^\epsilon, \mathbf{a}^\epsilon, \ldots)$, where*

$$\mathbf{a}^\epsilon(i) = \arg\max_{a \in \mathcal{A}} \left\{ c(i,a) + \nu \hat{\sigma}_i^a \right\}, \quad i \in \mathcal{X}.$$

In (Nilim & El-Ghaoui 2003; 2004), we establish that the above algorithm £nds an ϵ-suboptimal robust policy in at most $O(mn^2 \log(1/\epsilon)^2)$ ¤ops. Thus, the extra computational cost incurred by robustness in the in£nite-horizon case is only $O(\log(1/\epsilon))$.

Solving the inner problem

Each step of the robust dynamic programming algorithm involves the solution of an optimization problem, referred to as the "inner problem", of the form

$$\sigma_{\mathcal{P}}(v) = \max_{p \in \mathcal{P}} v^T p, \tag{15}$$

where the variable p corresponds to a particular row of a speci£c transition matrix, $\mathcal{P} = \mathcal{P}_i^a$ is the set that describes the uncertainty on this row, and v contains the elements of the value function at some given stage. The complexity of the sets $\mathcal{P}_i^a$ for each $i \in \mathcal{X}$ and $a \in \mathcal{A}$ is a key component in the complexity of the robust dynamic programming algorithm. Note that we can safely replace $\mathcal{P}$ in (15) by its convex hull, so that convexity of the sets $\mathcal{P}_i^a$ is not required; the algorithm only requires the knowledge of their convex hulls.

Beyond numerical tractability, an additional criteria for the choice of a speci£c uncertainty model is that the sets $\mathcal{P}^a$ should represent accurate (non-conservative) descriptions of the statistical uncertainty on the transition matrices. Perhaps surprisingly, there are statistical models of uncertainty described via Kullback-Liebler divergence that are good on both counts; speci£c examples of such models are described in the following sections.

Kullback-Liebler Divergence Uncertainty Models

We now address the inner problem (15) for a speci£c action $a \in \mathcal{A}$ and state $i \in \mathcal{X}$. Denote by $D(p\|q)$ denotes the Kullback-Leibler (KL) divergence (relative entropy) from the probability distribution $q \in \Delta_n$ to the probability distribution $p \in \Delta_n$:

$$D(p\|q) := \sum_j p(j) \log \frac{p(j)}{q(j)}.$$

The above function provides a natural way to describe errors in (rows of) the transition matrices; examples of models based on this function are given below.

Likelihood Models: Our £rst uncertainty model is derived from a controlled experiment starting from state $i = 1, 2, \ldots, n$ and the count of the number of transitions to different states. We denote by F^a the matrix of empirical frequencies of transition with control a in the experiment; denote by f_i^a its ith row. We have $F^a \geq 0$ and $F^a \mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ denotes the vector of ones. The "plug-in" estimate $\hat{P}^a = F^a$ is the solution to the maximum likelihood problem

$$\max_P \sum_{i,j} F^a(i,j) \log P(i,j) \ : \ P \geq 0, \ \ P\mathbf{1} = \mathbf{1}. \tag{16}$$

The optimal log-likelihood is $\beta_{\max}^a = \sum_{i,j} F^a(i,j) \log F^a(i,j)$. A classical description of uncertainty in a maximum-likelihood setting is via the "likelihood region" (Lehmann & Casella 1998) $\mathcal{P}^a = \{P \in \mathbf{R}^{n \times n} : P \geq 0, \ P\mathbf{1} = \mathbf{1}, \ \sum_{i,j} F^a(i,j) \log P(i,j) \geq \beta^a\}$,
where $\beta^a < \beta_{\max}^a$ is a pre-speci£ed number, which represents the uncertainty level. In practice, the designer speci£es an uncertainty level β^a based on re-sampling methods, or on a large-sample Gaussian approximation, so as to ensure that the set above achieves a desired level of con£dence.

With the above model, we note that the inner problem (15) only involves the set $\mathcal{P}_i^a := \left\{ p_i^a \in \mathbf{R}^n \ : \ p_i^a \geq 0, \ \ p_i^{a\,T} \mathbf{1} = 1, \ \ \sum_j F^a(i,j) \log p_i^a(j) \geq \beta_i^a \right\}$, where $\beta_i^a := \beta^a - \sum_{k \neq i} \sum_j F^a(k,j) \log F^a(k,j)$. The set

$\mathcal{P}_i^a$ is the *projection* of the set described above on a speci£c axis of p_i^a-variables. Noting further that the likelihood function can be expressed in terms of KL divergence, the corresponding uncertainty model on the row p_i^a for given $i \in \mathcal{X}$, $a \in \mathcal{A}$, is given by a set of the form $\mathcal{P}_i^a = \{p \in \Delta_n : D(f_i^a \| p) \leq \gamma_i^a\}$, where $\gamma_i^a = \sum_j F^a(i,j) \log F^a(i,j) - \beta_i^a$ is a function of the uncertainty level, and f_i^a is the ith row of the matrix F^a.

Maximum A-Posteriori (MAP) Models: a variation on Likelihood models involves Maximum A Posteriori (MAP) estimates. If there exist a prior information regarding the uncertainty on the i-th row of P^a, which can be described via a Dirichlet distribution (Ferguson 1974) with parameter α_i^a, the resulting MAP estimation problem takes the form

$$\max_p \, (f_i^a + \alpha_i^a - \mathbf{1})^T \log p \; : \; p^T \mathbf{1} = 1, p \geq 0.$$

Thus, the MAP uncertainty model is equivalent to a Likelihood model, with the sample distribution f_i^a replaced by $f_i^a + \alpha_i^a - \mathbf{1}$, where α_i^a is the prior corresponding to state i and action a.

Relative Entropy Models: Likelihood or MAP models involve the KL divergence from the unknown distribution to a reference distribution. We can also choose to describe uncertainty by exchanging the order of the arguments of the KL divergence. This results in a so-called "relative entropy" model, where the uncertainty on the i-th row of the transition matrix P^a described by a set of the form $\mathcal{P}_i^a = \{p \in \Delta_n : D(p \| q_i^a) \leq \gamma_i^a\}$, where $\gamma_i^a > 0$ is £xed, $q_i^a > 0$ is a given "reference" distribution (for example, the Maximum Likelihood distribution).

Inner problem with Likelihood Models

The inner problem (15) with the Likelihood uncertainty model is the following,

$$\sigma^* := \max_p \, p^T v \; : \; p \in \Delta^n, \; \sum_j f(j) \log p(j) \geq \beta, \quad (17)$$

where we have dropped the subscript i and superscript a in the empirical frequencies vector f_i^a and in the lower bound β_i^a. In this section $\beta_{\max}$ denotes the maximal value of the likelihood function appearing in the above set, which is $\beta_{\max} = \sum_j f(j) \log f(j)$. We assume that $\beta < \beta_{\max}$, which, together with $f > 0$, ensures that the set above has non-empty interior. Without loss of generality, we can assume that $v \in \mathbf{R}_+^n$.

The dual problem

The Lagrangian $\mathcal{L} : \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ associated with the inner problem can be written as

$$\mathcal{L}(v, \zeta, \mu, \lambda) = p^T v + \zeta^T p + \mu(1 - p^T \mathbf{1}) + \lambda(f^T \log p - \beta),$$

where ζ, μ, and λ are the Lagrange multipliers. The Lagrange dual function $d : \mathbf{R}^n \times \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ is the maximum value of the Lagrangian over p, i.e., for $\zeta \in \mathbf{R}^n$, $\mu \in \mathbf{R}$, and $\lambda \in \mathbf{R}$,

$$\begin{aligned} d(\zeta, \mu, \lambda) &= \sup_p \mathcal{L}(v, \zeta, \mu, \lambda) \quad (18) \\ &= \sup_p (p^T v + \zeta^T p + \mu(1 - p^T \mathbf{1}) \\ &\qquad + \lambda(f^T \log p - \beta)). \end{aligned}$$

The optimal $p^* = \arg\sup_p \mathcal{L}(v, \zeta, \mu, \lambda)$ is readily be obtained by solving $\frac{\partial \mathcal{L}}{\partial p} = 0$, which results in

$$p^*(i) = \frac{\lambda f(i)}{\mu - v(i) - \zeta(i)}.$$

Plugging the value of p^* in the equation for $d(\nu, \mu, \lambda)$ yields, with some simpli£cation, the following dual problem:

$$\overline{\sigma} := \min_{\lambda, \mu, \zeta} \; \mu - (1 + \beta)\lambda + \lambda \sum_j f(j) \log \frac{\lambda f(j)}{\mu - v(j) - \nu(j)},$$

$$\text{such that } \lambda \geq 0, \;\; \zeta \geq 0, \;\; \zeta + v \leq \mu \mathbf{1}.$$

Since the above problem is convex, and has a feasible set with non-empty interior, there is no duality gap, that is, $\sigma^* = \overline{\sigma}$. Moreover, by a monotonicity argument, we obtain that the optimal dual variable ζ is zero, which reduces the number of variables to two:

$$\sigma^* = \min_{\lambda, \mu} \; h(\lambda, \mu)$$

where

$$h(\lambda, \mu) := \begin{cases} \mu - (1 + \beta)\lambda + & \\ \lambda \sum_j f(j) \log \dfrac{\lambda f(j)}{\mu - v(j)} & \text{if } \lambda > 0, \;\; \mu > v_{\max}, \\ +\infty & \text{otherwise.} \end{cases} \quad (19)$$

(Note that, $v_{\max} := \max_j v(j)$).

For further reference, we note that h is twice differentiable on its domain, and that its gradient is given by

$$\nabla h(\lambda, \mu) = \begin{bmatrix} \sum_j f(j) \log \dfrac{\lambda f(j)}{\mu - v(j)} - \beta \\ 1 - \lambda \sum_j \dfrac{f(j)}{\mu - v(j)} \end{bmatrix}. \quad (20)$$

A bisection algorithm

From the expression of the gradient obtained above, we obtain that the optimal value of λ for a £xed μ, $\lambda(\mu)$, is given analytically by

$$\lambda(\mu) = \left(\sum_j \frac{f(j)}{\mu - v(j)} \right)^{-1}, \quad (21)$$

which further reduces the problem to a one-dimensional problem:

$$\sigma^* = \min_{\mu \geq v_{\max}} \; \sigma(\mu),$$

where $v_{\max} = \max_j v(j)$, and $\sigma(\mu) = h(\lambda(\mu), \mu)$. By construction, the function $\sigma(\mu)$ is convex in its (scalar) argument, since the function h de£ned in (19) is jointly convex

in both its arguments (see (Boyd & Vandenberghe January 2004, p.74)). Hence, we may use bisection to minimize σ.

To initialize the bisection algorithm, we need upper and lower bounds μ_- and μ_+ on a minimizer of σ. When $\mu \to v_{\max}$, $\sigma(\mu) \to v_{\max}$ and $\sigma'(\mu) \to -\infty$ (see (Nilim & El-Ghaoui 2004)). Thus, we may set the lower bound to $\mu_- = v_{\max}$.

The upper bound μ_+ must be chosen such that $\sigma'(\mu_+) > 0$. We have

$$\sigma'(\mu) = \frac{\partial h}{\partial \mu}(\lambda(\mu), \mu) + \frac{\partial h}{\partial \lambda}(\lambda(\mu), \mu)\frac{d\lambda(\mu)}{d\mu}. \qquad (22)$$

The £rst term is zero by construction, and $d\lambda(\mu)/d\mu > 0$ for $\mu > v_{\max}$. Hence, we only need a value of μ for which

$$\frac{\partial h}{\partial \lambda}(\lambda(\mu), \mu) = \sum_j f(j) \log \frac{\lambda(\mu) f(j)}{\mu - v(j)} - \beta > 0. \qquad (23)$$

By convexity of the negative log function, and using the fact that $f^T \mathbf{1} = 1$, $f \geq 0$, we obtain that

$$\begin{aligned} \frac{\partial h}{\partial \lambda}(\lambda(\mu), \mu) &= \beta_{\max} - \beta + \textstyle\sum_j f(j) \log \dfrac{\lambda(\mu)}{\mu - v(j)} \\ &\geq \beta_{\max} - \beta - \log \left(\textstyle\sum_j f(j) \dfrac{\mu - v(j)}{\lambda(\mu)} \right) \\ &\geq \beta_{\max} - \beta + \log \frac{\lambda(\mu)}{\mu - \bar{v}}, \end{aligned}$$

where $\bar{v} = f^T v$ denotes the average of v under f.

The above, combined with the bound on $\lambda(\mu)$: $\lambda(\mu) \geq \mu - v_{\max}$, yields a suf£cient condition for (23) to hold:

$$\mu > \mu_+^0 := \frac{v_{\max} - e^{\beta - \beta_{\max}} \bar{v}}{1 - e^{\beta - \beta_{\max}}}. \qquad (24)$$

By construction, the interval $[v_{\max}, \mu_+^0]$ is guaranteed to contain a global minimizer of σ over $(v_{\max}, +\infty)$.

The bisection algorithm goes as follows:

1. Set $\mu_- = v_{\max}$ and $\mu_+ = \mu_+^0$ as in (24). Let $\delta > 0$ be a small convergence parameter.
2. While $\mu_+ - \mu_- > \delta(1 + \mu_- + \mu_-)$, repeat
   (a) Set $\mu = (\mu_+ + \mu_-)/2$.
   (b) Compute the gradient of σ at μ.
   (c) If $\sigma'(\mu) > 0$, set $\mu_+ = \mu$; otherwise, set $\mu_- = \mu$.
   (d) go to 2a.

In practice, the function to minimize may be very "¤at" near the minimum. This means that the above bisection algorithm may take a long time to converge to the global minimizer. Since we are only interested in the value of the minimum (and not of the minimizer), we may modify the stopping criterion to

$$\mu_+ - \mu_- \leq \delta(1 + \mu_- + \mu_-) \text{ or } \sigma'(\mu_+) - \sigma'(\mu_-) \leq \delta.$$

The second condition in the criterion implies that $|\sigma'((\mu_+ + \mu_-)/2)| \leq \delta$, which is an approximate condition for global optimality.

Inner problem with Maximum A Posteriori Models

The inner problem (15) with MAP uncertainty models takes the form

$$\sigma^* := \max_p \; p^T v \; : \; p \geq 0, \;\; p^T \mathbf{1} = 1,$$
$$\sum_j \left(f(j) + \alpha(j) - 1\right) \log p(j) \geq \kappa,$$

where κ depends on the normalizing constant K appearing in the prior density function and on the chosen lower bound on the MAP function, β. We observe that this problem has exactly the same form as in the case of likelihood function, provided we replace f by $f + \alpha - \mathbf{1}$. Therefore, the same results apply to the MAP case.

Inner problem with Entropy Models

The inner problem (15) with entropy uncertainty models takes the form

$$\sigma^* := \max_p p^T v : \sum_j p(j) \log \frac{p(j)}{q(j)} \geq \beta \; : \; p^T \mathbf{1} = 1, p \geq 0.$$

We note that the constraint set actually equals the whole probability simplex if β is too large, speci£cally if $\beta \geq \max_i(-\log q_i)$, since the latter quantity is the maximum of the relative entropy function over the simplex. Thus, if $\beta \geq \max_i(-\log q_i)$, the worst-case value of $p^T v$ for $p \in \mathcal{P}$ is equal to $v_{\max} := \max_j v(j)$.

Dual problem

By standard duality arguments (set $\mathcal{P}$ being of non-empty interior), the inner problem is equivalent to its dual:

$$\min_{\lambda > 0, \mu} \; \mu + \beta\lambda + \lambda \sum_j q(j) \exp\left(\frac{v(j) - \mu}{\lambda} - 1\right).$$

Setting the derivative with respect to μ to zero, we obtain the optimality condition

$$\sum_j q(j) \exp\left(\frac{v(j) - \mu}{\lambda} - 1\right) = 1,$$

from which we derive

$$\mu = \lambda \log \left(\sum_j q(j) \exp \frac{v(j)}{\lambda} \right) - \lambda.$$

The optimal distribution is

$$p^* = \frac{q(j) \exp \frac{v(j)}{\lambda}}{\sum_i q(i) \exp \frac{v(i)}{\lambda}}.$$

As before, we reduce the problem to a one-dimensional problem:

$$\min_{\lambda > 0} \; \sigma(\lambda)$$

where σ is the convex function:

$$\sigma(\lambda) = \lambda \log \left(\sum_j q(j) \exp \frac{v(j)}{\lambda} \right) + \beta\lambda. \qquad (25)$$

Perhaps not surprisingly, the above function is closely linked to the moment generating function of a random variable $\mathbf{v}$ having the discrete distribution with mass q_i at v_i.

A bisection algorithm

As proved in (Nilim & El-Ghaoui 2004), the convex function σ in (25) has the following properties:

$$\forall\, \lambda \geq 0, \;\; q^T v + \beta\lambda \leq \sigma(\lambda) \leq v_{\max} + \beta\lambda, \qquad (26)$$

and

$$\sigma(\lambda) = v_{\max} + (\beta + \log Q(v))\lambda + o(\lambda), \qquad (27)$$

where

$$Q(v) := \sum_{j\,:\,v(j)=v_{\max}} q(j) = \mathbf{Prob}\{\mathbf{v} = v_{\max}\}.$$

Hence, $\sigma(0) = v_{\max}$ and $\sigma'(0) = \beta + \log Q(v)$. In addition, at in£nity the expansion of σ is

$$\sigma(\lambda) = q^T v + \beta\lambda + o(1). \qquad (28)$$

The bisection algorithm can be started with the lower bound $\lambda_- = 0$. An upper bound can be computed by £nding a solution to the equations $\sigma(0) = q^T v + \beta\lambda$, which yields the initial upper bound $\lambda_+^0 = (v_{\max} - q^T v)/\beta$. By convexity, a minimizer exists in the interval $[0\ \lambda_+^0]$.

Note that if $\sigma'(0) \geq 0$, then $\lambda = 0$ is optimal and the optimal value of σ is $v_{\max}$. This means that if β is too high, that is, if $\beta > -\log Q(v)$, enforcing robustness amounts to disregard any prior information on the probability distribution p. We have observed in (Nilim & El-Ghaoui 2004) a similar phenomenon brought about by too large values of β, which resulted in a set $\mathcal{P}$ equal to the probability simplex. Here, the limiting value $-\log Q(v)$ depends not only on q but also on v, since we are dealing with the optimization problem (15) and not only with its feasible set $\mathcal{P}$.

Computational complexity of the inner problem

Equipped with one of the above uncertainty models, we have shown in the previous section that the inner problem can be converted by convex duality, to a problem of minimizing a single-variable, convex function. In turn, this one-dimensional convex optimization problem can be solved via a bisection algorithm with a worst-case complexity of $O(n \log(v_{\max}/\delta))$ (see (Nilim & El-Ghaoui 2004) for details), where $\delta > 0$ speci£es the accuracy at which the optimal value of the inner problem (15) is computed, and $v_{\max}$ is a global upper bound on the value function.

Remark: We can also use models where the uncertainty in the i-th row for the transition matrix P^a is described by a £nite set of vectors, $\mathcal{P}_i^a = \{p_i^{a,1}, \ldots, p_i^{a,K}\}$. In this case the complexity of the corresponding robust dynamic programming algorithm is increased by a relative factor of K with respect to its classical counterpart, which makes the approach attractive when the number of "scenarios" K is moderate.

Concluding remarks

We proposed a "robust dynamic programming" algorithm for solving £nite-state and £nite-action MDPs whose solutions are guaranteed to tolerate arbitrary changes of the transition probability matrices within given sets. We proposed models based on KL divergence, which is a natural way to describe estimation errors. The resulting robust dynamic programming algorithm has almost the same computational cost as the classical dynamic programming algorithm: the relative increase to compute an ϵ-suboptimal policy is $O(\log(N/\epsilon))$ in the N-horizon case, and $O(\log(1/\epsilon))$ for the in£nite-horizon case.

References

Bagnell, J.; Ng, A.; and Schneider, J. 2001. Solving uncertain Markov decision problems. Technical Report CMU-RI-TR-01-25, Robotics Institute, Carnegie Mellon University.

Boyd, S., and Vandenberghe, L. January, 2004. *Convex Optimization*. Cambridge, U.K.: Cambridge University Press.

Feinberg, E., and Shwartz, A. 2002. *Handbook of Markov Decision Processes, Methods and Applications*. Boston: Kluwer's Academic Publishers.

Ferguson, T. 1974. Prior distributions on space of probability measures. *The Annal of Statistics* 2(4):615–629.

Givan, R.; Leach, S.; and Dean, T. 1997. Bounded Parameter Markov Decision Processes. In *fourth European Conference on Planning*, 234–246.

Lehmann, E., and Casella, G. 1998. *Theory of point estimation*. New York, USA: Springer-Verlag.

Nilim, A., and El-Ghaoui, L. 2003. Robust solution to Markov decision problems with uncertain transition matrices: proofs and complexity analysis. In *the proceeding of the Neural Information Processing Systems (NIPS, 2004)*.

Nilim, A., and El-Ghaoui, L. 2004. Curse of uncertainty in markaov decision processes: how tro £x it. Technical Report UCB/ERL M04/, Department of EECS, University of California, Berkeley. A related version has been submitted to *Operations Research* in Dec. 2003.

Puterman, M. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: Wiley-Interscience.

Satia, J. K., and Lave, R. L. 1973. Markov decision processes with uncertain transition probabilities. *Operations Research* 21(3):728–740.

Shapiro, A., and Kleywegt, A. J. 2002. Minimax analysis of stochastic problems. *Optimization Methods and Software*. to appear.

White, C. C., and Eldeib, H. K. 1994. Markov decision processes with imprecise transition probabilities. *Operations Research* 42(4):739–749.