Bounded Parameter Markov Decision Processes with Average Reward Criterion

Ambuj Tewari[1] and Peter L. Bartlett[2]

[1] University of California, Berkeley
Division of Computer Science
544 Soda Hall # 1776
Berkeley, CA 94720-1776, USA
ambuj@cs.berkeley.edu
[2] University of California, Berkeley
Division of Computer Science and Department of Statistics
387 Soda Hall # 1776
Berkeley, CA 94720-1776, USA
bartlett@cs.berkeley.edu

Abstract. Bounded parameter Markov Decision Processes (BMDPs) address the issue of dealing with uncertainty in the parameters of a Markov Decision Process (MDP). Unlike the case of an MDP, the notion of an optimal policy for a BMDP is not entirely straightforward. We consider two notions of optimality based on optimistic and pessimistic criteria. These have been analyzed for discounted BMDPs. Here we provide results for average reward BMDPs.

We establish a fundamental relationship between the discounted and the average reward problems, prove the existence of Blackwell optimal policies and, for both notions of optimality, derive algorithms that converge to the optimal value function.

1 Introduction

Markov Decision Processes (MDPs) are a widely used tool to model decision making under uncertainty. In an MDP, the uncertainty involved in the outcome of making a decision in a certain state is represented using various probabilities. However, these probabilities themselves may not be known precisely. This can happen for a variety of reasons. The probabilities might have been obtained via an estimation process. In such a case, it is natural that confidence intervals will be associated with them. State aggregation, where groups of similar states of a large MDP are merged to form a smaller MDP, can also lead to a situation where probabilities are no longer known precisely but are only known to lie in an interval.

This paper is concerned with such higher level uncertainty, namely uncertainty about the parameters of an MDP. Bounded parameter MDPs (BMDPs) have been introduced in the literature [1] to address this problem. They use intervals (or equivalently, lower and upper bounds) to represent the set in which the

N. Bshouty and C. Gentile (Eds.): COLT 2007, LNAI 4539, pp. 263–277, 2007.

parameters of an MDP can lie. We obtain an entire family, say $\mathcal{M}$, of MDPs by taking all possible choices of parameters consistent with these intervals. For an exact MDP M and a policy μ (which is a mapping specifying the actions to take in various states), the α-discounted return from state i, $V_{\alpha,\mu,M}(i)$ and the long term average return $V_{\mu,M}(i)$ are two standard ways of measuring the quality of μ with respect to M. When we have a family $\mathcal{M}$ of MDPs, we are immediately faced with the problem of finding a way to measure the quality of a policy. An optimal policy will then be the one that maximizes the particular performance measure chosen.

We might choose to put a distribution over $\mathcal{M}$ and define the return of a policy as its average return under this distribution. In this paper, however, we will avoid taking this approach. Instead, we will consider the worst and the best MDP for each policy and accordingly define two performance measures,

$$V_\mu^{\text{opt}}(i) := \sup_{M\in\mathcal{M}} V_{\mu,M}(i)$$
$$V_\mu^{\text{pes}}(i) := \inf_{M\in\mathcal{M}} V_{\mu,M}(i)$$

where the superscripts denote that these are optimistic and pessimistic criteria respectively. Analogous quantities for the discounted case were defined in [1] and algorithms were given to compute them. In this paper, our aim is to analyze the average reward setting.

The optimistic criterion is motivated by the *optimism in the face of uncertainty* principle. Several learning algorithms for MDPs [2,3,4,5] proceed in the following manner. Faced with an unknown MDP, they start collecting data which yields confidence intervals for the parameters of the MDP. Then they choose a policy which is optimal in the sense of the optimistic criterion. This policy is followed for the next phase of data collection and the process repeats. In fact, the algorithm of Auer and Ortner requires, as a blackbox, an algorithm to compute the optimal (with respect to the optimistic criterion) value function for a BMDP.

The pessimistic criterion is related to research on robust control of MDPs [6]. If nature is adversarial, then once we pick a policy μ it will pick the worst possible MDP M from $\mathcal{M}$. In such a scenario, it is reasonable to choose a policy which is best in the worst case. Our work also extends this line of research to the case of the average reward criterion.

A brief outline of the paper is as follows. Notation and preliminary results are established in Section 2. Most of these results are not new but are needed later, and we provide independent, self-contained proofs in the appendix. Section 3 proves one of the key results of the paper: the existence of Blackwell optimal policies. In the exact MDP case, a Blackwell optimal policy is a policy that is optimal for an entire range of discount factors in the neighbourhood of 1. Existence of Blackwell optimal policies is an important result in the theory of MDPs. We extend this result to BMDPs. Then, in Section 4, we exploit the relationship between the discounted and average returns together with the

existence of a Blackwell optimal policy to derive algorithms that converge to optimal value functions for both optimistic as well as pessimistic criteria.

2 Preliminaries

A Markov Decision Process is a tuple $\langle S, A, R, \{p_{(i,j)}(a)\}\rangle$. Here S is a finite set of states, A a finite set of actions, $R : S \mapsto [0,1]$ is the reward function and $p_{i,j}(a)$ is the probability of moving to state j upon taking action a in state i. A policy $\mu : S \mapsto A$ is a mapping from states to actions. Any policy induces a Markov chain on the state space of a given MDP M. Let $\mathbb{E}_{\mu,M}[\cdot]$ denote expectation taken with respect to this Markov chain. For $\alpha \in [0,1)$, define the α-discounted value function at state $i \in S$ by

$$V_{\alpha,\mu,M}(i) := (1-\alpha)\mathbb{E}_{\mu,M}\left[\sum_{t=0}^{\infty} \alpha^t R(s_t) \,\middle|\, s_0 = i\right] .$$

The optimal value function is obtained by maximizing over policies.

$$V^*_{\alpha,M}(i) := \max_{\mu} V_{\alpha,\mu,M}(i) .$$

From the definition it is not obvious that there is a single policy achieving the maximum above for all $i \in S$. However, it is a fundamental result of the theory of MDPs that such an optimal policy exists.

Instead of considering the discounted sum, we can also consider the long term average reward. This leads us to the following definition.

$$V_{\mu,M}(i) := \lim_{T\to\infty} \frac{\mathbb{E}_{\mu,M}\left[\sum_{t=0}^{T} R(s_t) \,\middle|\, s_0 = i\right]}{T+1}$$

The above definition assumes that the limit on the right hand side exists for every policy. This is shown in several standard texts [7]. There is an important relationship between the discounted and undiscounted value functions of a policy. For every policy μ, there is a function $h_{\mu,M} : S \mapsto \mathbb{R}$ such that

$$\forall i,\ V_{\mu,M}(i) = V_{\alpha,\mu,M}(i) + (1-\alpha)h_{\mu,M}(i) + O\left(|1-\alpha|^2\right) . \tag{1}$$

A bounded parameter MDP (BMDP) is a collection of MDPs specified by bounds on the parameters of the MDPs. For simplicity, we will assume that the reward function is fixed, so that the only parameters that vary are the transition probabilities. Suppose, for each state-action pair i, a, we are given lower and upper bounds, $l(i,j,a)$ and $u(i,j,a)$ respectively, on the transition probability $p_{i,j}(a)$. We assume that the bounds are legitimate, that is

$$\forall i,a,j,\ 0 \le l(i,j,a) \le u(i,j,a) ,$$

$$\forall i,a,\ \sum_j l(i,j,a) \le 1 \ \&\ \sum_j u(i,j,a) \ge 1 .$$

This means that the set defined[1] by

$$\mathcal{C}_{i,a} := \{q \in \mathbb{R}_+^{|S|} \ : \ q^T \mathbf{1} = 1 \ \& \ \forall j, \ l(i,j,a) \leq q_j \leq u(i,j,a)\}$$

is non-empty for each state-action pair i, a. Finally, define the collection of MDPs

$$\mathcal{M} := \{ \ \langle S, A, R, \{p_{i,j}(a)\}\rangle \ : \ \forall i, a, \ p_{i,\cdot}(a) \in \mathcal{C}_{i,a} \ \} \ .$$

Given a BMDP $\mathcal{M}$ and a policy μ, there are two natural choices for the value function: an optimistic and a pessimistic one,

$$V_{\alpha,\mu}^{\text{opt}}(i) := \sup_{M \in \mathcal{M}} V_{\alpha,\mu,M}(i) \qquad\qquad V_{\alpha,\mu}^{\text{pes}}(i) := \inf_{M \in \mathcal{M}} V_{\alpha,\mu,M}(i) \ .$$

We also define the undiscounted value functions,

$$V_{\mu}^{\text{opt}}(i) := \sup_{M \in \mathcal{M}} V_{\mu,M}(i) \qquad\qquad V_{\mu}^{\text{pes}}(i) := \inf_{M \in \mathcal{M}} V_{\mu,M}(i) \ .$$

Optimal value functions are defined by maximizing over policies.

$$\mathbf{V}_{\alpha}^{\text{opt}}(i) := \max_{\mu} V_{\alpha,\mu}^{\text{opt}}(i) \qquad\qquad \mathbf{V}_{\alpha}^{\text{pes}}(i) := \max_{\mu} V_{\alpha,\mu}^{\text{pes}}(i)$$
$$\mathbf{V}^{\text{opt}}(i) := \max_{\mu} V_{\mu}^{\text{opt}}(i) \qquad\qquad \mathbf{V}^{\text{pes}}(i) := \max_{\mu} V_{\mu}^{\text{pes}}(i)$$

In this paper, we are interested in computing $\mathbf{V}^{\text{opt}}$ and $\mathbf{V}^{\text{pes}}$. Algorithms to compute $\mathbf{V}_{\alpha}^{\text{opt}}$ and $\mathbf{V}_{\alpha}^{\text{pes}}$ have already been proposed in the literature. Let us review some of the results pertaining to the discounted case. We note that the results in this section, with the exception of Corollary 4, either appear or can easily be deduced from results appearing in [1]. However, we provide self-contained proofs of these in the appendix. Before we state the results, we need to introduce a few important operators. Note that, since $\mathcal{C}_{i,a}$ is a closed, convex set, the maximum (or minimum) of $q^T V$ (a linear function of q) appearing in the definitions below is achieved.

$$\left(T_{\alpha,\mu,M} V\right)(i) := (1-\alpha) R(i) + \alpha \sum_j p_{i,j}(\mu(i)) V(j)$$

$$\left(T_{\alpha,M} V\right)(i) := \max_{a \in A} \left[(1-\alpha) R(i) + \alpha \sum_j p_{i,j}(a) V(j) \right]$$

$$\left(T_{\alpha,\mu}^{\text{opt}} V\right)(i) := (1-\alpha) R(i) + \alpha \max_{q \in \mathcal{C}_{i,\mu(i)}} q^T V$$

$$\left(T_{\alpha}^{\text{opt}} V\right)(i) := \max_{a \in A} \left[(1-\alpha) R(i) + \alpha \max_{q \in \mathcal{C}_{i,a}} q^T V \right]$$

$$\left(T_{\alpha,\mu}^{\text{pes}} V\right)(i) := (1-\alpha) R(i) + \alpha \min_{q \in \mathcal{C}_{i,\mu(i)}} q^T V$$

$$\left(T_{\alpha}^{\text{pes}} V\right)(i) := \max_{a \in A} \left[(1-\alpha) R(i) + \alpha \min_{q \in \mathcal{C}_{i,a}} q^T V \right]$$

[1] We denote the transpose of a vector q by q^T.

Recall that an operator T is a contraction mapping with respect to a norm $\|\cdot\|$ if there is an $\alpha \in [0,1)$ such that

$$\forall V_1, V_2,\ \|TV_1 - TV_2\| \leq \alpha \|V_1 - V_2\| \ .$$

A contraction mapping has a unique solution to the fixed point equation $TV = V$ and the sequence $\{T^k V_0\}$ converges to that solution for any choice of V_0. It is straightforward to verify that the six operators defined above are contraction mappings (with factor α) with respect to the norm

$$\|V\|_\infty := \max_i |V(i)| \ .$$

It is well known that the fixed points of $T_{\alpha,\mu,M}$ and $T_{\alpha,M}$ are $V_{\alpha,\mu,M}$ and $V^*_{\alpha,M}$ respectively. The following theorem tells us what the fixed points of the remaining four operators are.

Theorem 1. *The fixed points of* $T^{\rm opt}_{\alpha,\mu}, T^{\rm opt}_{\alpha}, T^{\rm pes}_{\alpha,\mu}$ *and* $T^{\rm pes}_{\alpha}$ *are* $V^{\rm opt}_{\alpha,\mu}, \mathbf{V}^{\rm opt}_{\alpha}, V^{\rm pes}_{\alpha,\mu}$ *and* $\mathbf{V}^{\rm pes}_{\alpha}$ *respectively.*

Existence of optimal policies for BMDPs is established by the following theorem.

Theorem 2. *For any* $\alpha \in [0,1)$*, there exist optimal policies* μ_1 *and* μ_2 *such that, for all* $i \in S$*,*

$$V^{\rm opt}_{\alpha,\mu_1}(i) = \mathbf{V}^{\rm opt}_{\alpha}(i) \ ,$$
$$V^{\rm pes}_{\alpha,\mu_2}(i) = \mathbf{V}^{\rm pes}_{\alpha}(i) \ .$$

A very important fact is that out of the uncountably infinite set $\mathcal{M}$, only a finite set is of real interest.

Theorem 3. *There exist finite subsets* $\mathcal{M}_{\rm opt}, \mathcal{M}_{\rm pes} \subset \mathcal{M}$ *with the following property. For all* $\alpha \in [0,1)$ *and for every policy* μ *there exist* $M_1 \in \mathcal{M}_{\rm opt}$*,* $M_2 \in \mathcal{M}_{\rm pes}$ *such that*

$$V^{\rm opt}_{\alpha,\mu} = V_{\alpha,\mu,M_1} \ ,$$
$$V^{\rm pes}_{\alpha,\mu} = V_{\alpha,\mu,M_2} \ .$$

Corollary 4. *The optimal undiscounted value functions are limits of the optimal discounted value functions. That is, for all* $i \in S$*, we have*

$$\lim_{\alpha\to 1} \mathbf{V}^{\rm opt}_{\alpha}(i) = \mathbf{V}^{\rm opt}(i) \ , \tag{2}$$
$$\lim_{\alpha\to 1} \mathbf{V}^{\rm pes}_{\alpha}(i) = \mathbf{V}^{\rm pes}(i) \ . \tag{3}$$

Proof. Fix $i \in S$. We first prove (2). Using Theorem 3, we have

$$\mathbf{V}^{\rm opt}_{\alpha}(i) = \max_{\mu} \max_{M\in\mathcal{M}_{\rm opt}} V_{\alpha,\mu,M}(i) \ .$$

Therefore,

$$\begin{aligned}\lim_{\alpha\to 1}\mathbf{V}_\alpha^{\mathrm{opt}}(i) &= \lim_{\alpha\to 1}\max_{\mu}\max_{M\in\mathcal{M}_{\mathrm{opt}}} V_{\alpha,\mu,M}(i)\\ &= \max_{\mu}\max_{M\in\mathcal{M}_{\mathrm{opt}}}\lim_{\alpha\to 1} V_{\alpha,\mu,M}(i)\\ &= \max_{\mu}\max_{M\in\mathcal{M}_{\mathrm{opt}}} V_{\mu,M}(i)\\ &= \mathbf{V}^{\mathrm{opt}}(i)\ .\end{aligned}$$

The second equality holds because lim and max over a finite set commute. Note that finiteness is crucial here since lim and sup do not commute. The third equality follows from (1).

To prove (3), one repeats the steps above with appropriate changes. In this case, one additionally uses the fact that lim and min over a finite set also commute.

3 Existence of Blackwell Optimal Policies

Theorem 5. *There exist $\alpha_{\mathrm{opt}} \in (0,1)$, a policy μ_{opt} and an MDP $M_{\mathrm{opt}} \in \mathcal{M}_{\mathrm{opt}}$ such that*

$$\forall\alpha \in (\alpha_{\mathrm{opt}},1),\ V_{\alpha,\mu_{\mathrm{opt}},M_{\mathrm{opt}}} = \mathbf{V}_\alpha^{\mathrm{opt}}\ .$$

Similarly, there exist $\alpha_{\mathrm{pes}} \in (0,1)$, a policy μ_{pes} and an MDP $M_{\mathrm{pes}} \in \mathcal{M}_{\mathrm{pes}}$ such that

$$\forall\alpha \in (\alpha_{\mathrm{pes}},1),\ V_{\alpha,\mu_{\mathrm{pes}},M_{\mathrm{pes}}} = \mathbf{V}_\alpha^{\mathrm{pes}}\ .$$

Proof. Given an MDP $M = \langle S, A, R, \{p_{i,j}(a)\}\rangle$ and a policy μ, define the associated matrix P_μ^M by

$$P_\mu^M(i,j) := p_{i,j}(\mu(i))\ .$$

The value function $V_{\alpha,\mu,M}$ has a closed form expression.

$$V_{\alpha,\mu,M} = (1-\alpha)\left(I - \alpha P_\mu^M\right)^{-1} R$$

Therefore, for all i, the map $\alpha \mapsto V_{\alpha,\mu,M}(i)$ is a rational function of α. Two rational functions are either identical or intersect each other at a finite number of points. Further, the number of policies and the number of MDPs in $\mathcal{M}_{\mathrm{opt}}$ is finite. Therefore, for each i, there exists $\alpha_i \in [0,1)$ such that no two functions in the set

$$\{\alpha \mapsto V_{\alpha,\mu,M}(i)\ :\ \mu : S \mapsto A,\ M \in \mathcal{M}_{\mathrm{opt}}\}$$

intersect each other in the interval $(\alpha_i, 1)$. Let $\alpha_{\mathrm{opt}} = \max_i \alpha_i$. By Theorem 2, there is an optimal policy, say μ_{opt}, such that

$$V^{\mathrm{opt}}_{\alpha_{\mathrm{opt}},\mu_{\mathrm{opt}}} = \mathbf{V}^{\mathrm{opt}}_{\alpha_{\mathrm{opt}}}\ .$$

By Theorem 3, there is an MDP, say M_{opt}, in $\mathcal{M}_{\mathrm{opt}}$ such that

$$V_{\alpha_{\mathrm{opt}},\mu_{\mathrm{opt}},M_{\mathrm{opt}}} = V^{\mathrm{opt}}_{\alpha_{\mathrm{opt}},\mu_{\mathrm{opt}}} = \mathbf{V}^{\mathrm{opt}}_{\alpha_{\mathrm{opt}}}\ . \tag{4}$$

We now claim that

$$V_{\alpha,\mu_{\rm opt},M_{\rm opt}} = \mathbf{V}^{\rm opt}_{\alpha_{\rm opt}}$$

for all $\alpha \in (\alpha_{\rm opt}, 1)$. If not, there is an $\alpha' \in (\alpha_{\rm opt}, 1)$, a policy μ' and an MDP $M' \in \mathcal{M}_{\rm opt}$ such that

$$V_{\alpha',\mu_{\rm opt},M_{\rm opt}}(i) < V_{\alpha',\mu',M'}(i)$$

for some i. But this yields a contradiction, since (4) holds and by definition of $\alpha_{\rm opt}$, the functions

$$\alpha \mapsto V_{\alpha,\mu_{\rm opt},M_{\rm opt}}(i)$$

and

$$\alpha \mapsto V_{\alpha,\mu',M'}(i)$$

cannot intersect in $(\alpha_{\rm opt}, 1)$.

The proof of the existence of $\alpha_{\rm pes}, \mu_{\rm pes}$ and $M_{\rm pes}$ is based on similar arguments.

4 Algorithms to Compute the Optimal Value Functions

4.1 Optimistic Value Function

The idea behind our algorithm (Algorithm 1) is to start with some initial vector and perform a sequence of updates while increasing the discount factor at a certain rate. The following theorem guarantees that the sequence of value functions thus generated converge to the optimal value function. Note that if we held the discount factor constant at some value, say α, the sequence would converge to $\mathbf{V}^{\rm opt}_{\alpha}$.

Algorithm 1. Algorithm to Compute $\mathbf{V}^{\rm opt}$

$V^{(0)} \leftarrow \mathbf{0}$
for $k = 0, 1, \ldots$ **do**
  $\alpha_k \leftarrow \frac{k+1}{k+2}$
  for all $i \in S$ **do**
    $V^{(k+1)}(i) \leftarrow \max_{a \in A} \left[(1-\alpha_k) R(i) + \alpha_k \max_{q \in \mathcal{C}_{i,a}} q^T V^{(k)} \right]$
  end for
end for

Theorem 6. *Let $\{V^{(k)}\}$ be the sequence of functions generated by Algorithm 1. Then we have, for all $i \in S$,*

$$\lim_{k \to \infty} V^{(k)}(i) = \mathbf{V}^{\rm opt}(i) \ .$$

We need a few intermediate results before proving this theorem. Let α_{opt}, μ_{opt} and M_{opt} be as given by Theorem 5. To avoid too many subscripts, let μ and M denote μ_{opt} and M_{opt} respectively for the remainder of this subsection. From (1), we have that for k large enough, say $k \geq k_1$, we have,

$$\left|V_{\alpha_k,\mu,M}(i) - V_{\alpha_{k+1},\mu,M}(i)\right| \leq K(\alpha_{k+1} - \alpha_k) \ , \tag{5}$$

where K can be taken to be $\|h_{\mu,M}\|_\infty + 1$. Since $\alpha_k \uparrow 1$, we have $\alpha_k > \alpha_{\text{opt}}$ for all $k > k_2$ for some k_2. Let $k_0 = \max\{k_1, k_2\}$. Define

$$\delta_{k_0} := \|V^{(k_0)} - V_{\alpha_{k_0},\mu,M}\|_\infty \ . \tag{6}$$

Since rewards are in $[0,1]$, we have $\delta_{k_0} \leq 1$. For $k \geq k_0$, define δ_{k+1} recursively as

$$\delta_{k+1} := K(\alpha_{k+1} - \alpha_k) + \alpha_k \delta_k \ . \tag{7}$$

The following lemma shows that this sequence bounds the norm of the difference between $V^{(k)}$ and $V_{\alpha_k,\mu,M}$.

Lemma 7. *Let $\{V^{(k)}\}$ be the sequence of functions generated by Algorithm 1. Further, let μ, M denote $\mu_{\text{opt}}, M_{\text{opt}}$ mentioned in Theorem 5. Then, for $k \geq k_0$, we have*

$$\|V^{(k)} - V_{\alpha_k,\mu,M}\|_\infty \leq \delta_k \ .$$

Proof. Base case of $k = k_0$ is true by definition of δ_{k_0}. Now assume we have proved the claim till $k \geq k_0$. So we know that,

$$\max_i \left|V^{(k)}(i) - V_{\alpha_k,\mu,M}(i)\right| \leq \delta_k \ . \tag{8}$$

We wish to show

$$\max_i \left|V^{(k+1)}(i) - V_{\alpha_{k+1},\mu,M}(i)\right| \leq \delta_{k+1} \ . \tag{9}$$

Recall that $\mathbf{V}_\alpha^{\text{opt}}$ is the fixed point of T_α^{opt} by Theorem 1. We therefore have, for all i,

$$\begin{aligned}
V_{\alpha_k,\mu,M}(i) &= \left(T_{\alpha_k}^{\text{opt}} V_{\alpha_k,\mu,M}\right)(i) \\
&\quad [\ \alpha_k > \alpha_{\text{opt}} \text{ and } V_{\alpha,\mu,M} = \mathbf{V}_\alpha^{\text{opt}} \text{ for } \alpha > \alpha_{\text{opt}} \ ] \\
&= \max_{a \in A}[\ (1-\alpha_k)R(i) + \alpha_k \max_{q \in \mathcal{C}_{i,a}} \sum_j q(j) V_{\alpha_k,\mu,M}(j) \ ] \\
&\quad [\ \text{defn. of } T_{\alpha_k}^{\text{opt}}] \\
&\leq \max_{a \in A}[\ (1-\alpha_k)R(i) + \alpha_k \max_{q \in \mathcal{C}_{i,a}} \sum_j q(j) V^{(k)}(j) \ ] + \alpha_k \delta_k \\
&\quad [\ (8) \text{ and } \sum_j q(j)\delta_k = \delta_k] \\
&= V^{(k+1)}(i) + \alpha_k \delta_k \ . \\
&\quad [\ \text{defn. of } V^{(k+1)}(i) \ ]
\end{aligned}$$

Similarly, for all i,

$$\begin{aligned}
V^{(k+1)}(i) &= \max_{a\in A}[\ (1-\alpha_k)R(i) + \alpha_k \max_{q\in\mathcal{C}_{i,a}} \sum_j q(j)V^{(k)}(j)\ ] \\
&\quad [\text{ defn. of } V^{(k+1)}(i)] \\
&\le \max_{a\in A}[\ (1-\alpha_k)R(i) + \alpha_k \max_{q\in\mathcal{C}_{i,a}} \sum_j q(j)V_{\alpha_k,\mu,M}(j)\ ] + \alpha_k\delta_k \\
&\quad [\text{ (8) and } \sum_j q(j)\delta_k = \delta_k] \\
&= \left(T^{\text{opt}}_{\alpha_k} V_{\alpha_k,\mu,M}\right)(i) + \alpha_k\delta_k \\
&\quad [\text{ defn. of } T^{\text{opt}}_{\alpha_k}] \\
&= V_{\alpha_k,\mu,M}(i) + \alpha_k\delta_k\ . \\
&\quad [\ \alpha_k > \alpha_{\text{opt}} \text{ and } V_{\alpha,\mu,M} = \mathbf{V}^{\text{opt}}_{\alpha} \text{ for } \alpha > \alpha_{\text{opt}}\ ]
\end{aligned}$$

Thus, for all i,

$$\left|V^{(k+1)}(i) - V_{\alpha_k,\mu,M}(i)\right| \le \alpha_k\delta_k\ .$$

Combining this with (5) (as $k \ge k_0 \ge k_1$), we get

$$\left|V^{(k+1)}(i) - V_{\alpha_{k+1},\mu,M}(i)\right| \le \alpha_k\delta_k + K(\alpha_{k+1} - \alpha_k)\ .$$

Thus we have shown (9).

The sequence $\{\delta_k\}$ can be shown to converge to zero using elementary arguments.

Lemma 8. *The sequence* $\{\delta_k\}$ *defined for* $k \ge k_0$ *by equations* (6) *and* (7) *converges to* 0.

Proof. Plugging $\alpha_k = \frac{k+1}{k+2}$ into the definition of δ_{k+1} we get,

$$\begin{aligned}
\delta_{k+1} &= K\left(\frac{k+2}{k+3} - \frac{k+1}{k+2}\right) + \frac{k+1}{k+2}\delta_k \\
&= \frac{K}{(k+3)(k+2)} + \frac{k+1}{k+2}\delta_k\ .
\end{aligned}$$

Applying the recursion again for δ_k, we get

$$\begin{aligned}
\delta_{k+1} &= \frac{K}{(k+3)(k+2)} + \frac{k+1}{k+2}\left(\frac{K}{(k+2)(k+1)} + \frac{k}{k+1}\delta_{k-1}\right) \\
&= \frac{K}{k+2}\left(\frac{1}{k+3} + \frac{1}{k+2}\right) + \frac{k}{k+2}\delta_{k-1}\ .
\end{aligned}$$

Continuing in this fashion, we get for any $j \ge 0$,

$$\delta_{k+1} = \frac{K}{k+2}\left(\frac{1}{k+3} + \frac{1}{k+2} + \ldots + \frac{1}{k-j+3}\right) + \frac{k-j+1}{k+2}\delta_{k-j}\ .$$

Setting $j = k - k_0$ above, we get

$$\delta_{k+1} = \frac{K}{k+2}\left(H_{k+3} - H_{k_0+2}\right) + \frac{k_0+1}{k+2}\delta_{k_0} \ ,$$

where $H_n = 1+\frac{1}{2}+\ldots+\frac{1}{n}$. This clearly tends to 0 as $k \to \infty$ since $H_n = O(\log n)$ and $\delta_{k_0} \leq 1$.

We can now prove Theorem 6.

Proof. (*of Theorem 6*) Fix $i \in S$. We have,

$$|V^{(k)}(i) - \mathbf{V}^{\mathrm{opt}}(i)| \leq \underbrace{|V^{(k)}(i) - V_{\alpha_k,\mu,M}(i)|}_{\leq \delta_k} + \underbrace{|V_{\alpha_k,\mu,M}(i) - \mathbf{V}^{\mathrm{opt}}_{\alpha_k}(i)|}_{\epsilon_k} + \underbrace{|\mathbf{V}^{\mathrm{opt}}_{\alpha_k}(i) - \mathbf{V}^{\mathrm{opt}}(i)|}_{\zeta_k} \ .$$

We use Lemma 7 to bound the first summand on the right hand side by δ_k. By Lemma 8, $\delta_k \to 0$. Also, $\epsilon_k = 0$ for sufficiently large k because $\alpha_k \uparrow 1$ and $V_{\alpha,\mu,M}(i) = \mathbf{V}^{\mathrm{opt}}_{\alpha}(i)$ for α sufficiently close to 1 (by Theorem 5). Finally, $\zeta_k \to 0$ by Corollary 4.

4.2 Pessimistic Value Function

Algorithm 2 is the same as Algorithm 1 except that the max over $\mathcal{C}_{i,a}$ appearing inside the innermost loop gets replaced by a min. The following analogue of Theorem 6 holds.

Algorithm 2. Algorithm to Compute $\mathbf{V}^{\mathrm{pes}}$

$V^{(0)} \leftarrow \mathbf{0}$
for $k = 0, 1, \ldots$ **do**
  $\alpha_k \leftarrow \frac{k+1}{k+2}$
  for all $i \in S$ **do**
    $V^{(k+1)}(i) \leftarrow \max_{a \in A}\left[(1-\alpha_k)R(i) + \alpha_k \min_{q \in \mathcal{C}_{i,a}} q^T V^{(k)}\right]$
  end for
end for

Theorem 9. *Let $\{V^{(k)}\}$ be the sequence of functions generated by Algorithm 2. Then we have, for all $i \in S$,*

$$\lim_{k\to\infty} V^{(k)}(i) = \mathbf{V}^{\mathrm{pes}}(i) \ .$$

To prove this theorem, we repeat the argument given in the previous subsection with appropriate changes. Let α_{pes}, μ_{pes} and M_{pes} be as given by Theorem 5. For the remainder of this subsection, let μ and M denote μ_{pes} and M_{pes} respectively. Let k_1, k_2 be large enough so that, for all $k \geq k_1$,

$$\left|V_{\alpha_k,\mu,M}(i) - V_{\alpha_{k+1},\mu,M}(i)\right| \leq K(\alpha_{k+1} - \alpha_k) \ ,$$

for some constant K (which depends on μ, M), and $\alpha_k > \alpha_{\text{pes}}$ for $k > k_2$. Set $k_0 = \max\{k_1, k_2\}$ and define the sequence $\{\delta_k\}_{k \geq k_0}$ as before (equations (6) and (7)).

The proof of the following lemma can be obtained from that of Lemma 7 by fairly straightforward changes and is therefore omitted.

Lemma 10. *Let $\{V^{(k)}\}$ be the sequence of functions generated by Algorithm 2. Further, let μ, M denote $\mu_{\text{pes}}, M_{\text{pes}}$ mentioned in Theorem 5. Then, for $k \geq k_0$, we have*

$$\|V^{(k)} - V_{\alpha_k,\mu,M}\|_\infty \leq \delta_k \ .$$

Theorem 9 is now proved in exactly the same fashion as Theorem 6 and we therefore omit the proof.

5 Conclusion

In this paper, we chose to represent the uncertainty in the parameters of an MDP by intervals. One can ask whether similar results can be derived for other representations. If the intervals for $p_{i,j}(a)$ are equal for all j then our representation corresponds to an L_∞ ball around a probability vector. It will be interesting to investigate other metrics and even non-metrics like relative entropy (for an example of an algorithm using sets defined by relative entropy, see [8]). Generalizing in a different direction, we can enrich the language used to express constraints on the probabilities. In this paper, constraints had the form

$$l(i,j,a) \leq p_{i,j}(a) \leq u(i,j,a) \ .$$

These are simple inequality constraints with two hyperparameters $l(i,j,a)$ and $u(i,j,a)$. We can permit more hyperparameters and include arbitrary semi-algebraic constraints (i.e. constraints expressible as boolean combination of polynomial equalities and inequalities). It can be shown using the Tarski-Seidenberg theorem that Blackwell optimal policies still exist in this much more general setting. However, the problem of optimizing $q^T V$ over $C_{i,a}$ now becomes more complicated.

Our last remark is regarding the convergence rate of the algorithms given in Section 4. Examining the proofs, one can verify that the number of iterations required to get to within ϵ accuracy is $O(\frac{1}{\epsilon})$. This is a pseudo-polynomial convergence rate. It might be possible to obtain algorithms where the number of iterations required to achieve ϵ-accuracy is poly$(\log \frac{1}{\epsilon})$.

Acknowledgments

We gratefully acknowledge the support of DARPA under grant FA8750-05-2-0249.

References

1. Givan, R., Leach, S., Dean, T.: Bounded-parameter Markov decision processes. Artificial Intelligence 122, 71–109 (2000)
2. Strehl, A.L., Littman, M.: A theoretical analysis of model-based interval estimation. In: Proceedings of the Twenty-Second International Conference on Machine Learning, pp. 857–864. ACM Press, New York (2005)
3. Auer, P., Ortner, R.: Logarithmic online regret bounds for undiscounted reinforcement learning. In: dvances in Neural Information Processing Systems 19, MIT Press, Cambridge (2007) (to appear)
4. Brafman, R.I., Tennenholtz, M.: R-MAX – a general polynomial time algorithm for near-optimal reinforcement learning. Journal of Machine Learning Research 3, 213–231 (2002)
5. Even-Dar, E., Mansour, Y.: Convergence of optimistic and incremental Q-learning. In: Advances in Neural Information Processing Systems 14, pp. 1499–1506. MIT Press, Cambridge (2001)
6. Nilim, A., El Ghaoui, L.: Robust control of Markov decision processes with uncertain transition matrices. Operations Research 53, 780–798 (2005)
7. Bertsekas, D.P.: Dynamic Programming and Optimal Control. Vol. 2. Athena Scientific, Belmont, MA (1995)
8. Burnetas, A.N., Katehakis, M.N.: Optimal adaptive policies for Markov decision processes. Mathematics of Operations Research 22, 222–255 (1997)

Appendix

Throughout this section, vector inequalities of the form $V_1 \leq V_2$ are to be interpreted to mean $V_1(i) \leq V_2(i)$ for all i.

Proofs of Theorems 1 and 2

Lemma 11. *If $V_1 \leq V_2$ then, for all $M \in \mathcal{M}$,*

$$T_{\alpha,\mu,M} V_1 \leq T^{\text{opt}}_{\alpha,\mu} V_2 \ ,$$
$$T^{\text{pes}}_{\alpha,\mu} V_1 \leq T_{\alpha,\mu,M} V_2 \ .$$

Proof. We prove the first inequality. Fix an MDP $M \in \mathcal{M}$. Let $p_{i,j}(a)$ denote transition probabilities of M. We then have,

$$\begin{aligned}
\left(T_{\alpha,\mu,M} V_1\right)(i) &= (1-\alpha) R(i) + \alpha \sum_j p_{i,j}(\mu(i)) V_1(j) \\
&\leq (1-\alpha) R(i) + \alpha \sum_j p_{i,j}(\mu(i)) V_2(j) && [\ \because V_1 \leq V_2\ ] \\
&\leq (1-\alpha) R(i) + \alpha \max_{q \in C_{i,\mu(i)}} q^T V_2 && [\ \because M \in \mathcal{M}\ ] \\
&= \left(T^{\text{opt}}_{\alpha,\mu} V_2\right)(i) \ .
\end{aligned}$$

The proof of the second inequality is similar.

Lemma 12. *If $V_1 \leq V_2$ then, for any policy μ,*

$$T^{\rm opt}_{\alpha,\mu} V_1 \leq T^{\rm opt}_{\alpha} V_2 \ ,$$
$$T^{\rm pes}_{\alpha,\mu} V_1 \leq T^{\rm pes}_{\alpha} V_2 \ .$$

Proof. Again, we prove only the first inequality. Fix a policy μ. We then have,

$$\begin{aligned}
\left(T^{\rm opt}_{\alpha,\mu} V_1\right)(i) &= (1-\alpha)R(i) + \alpha \max_{q \in \mathcal{C}_{i,\mu(i)}} q^T V_1 \\
&\leq (1-\alpha)R(i) + \alpha \max_{q \in \mathcal{C}_{i,\mu(i)}} q^T V_2 \\
&\leq \max_{a \in A} \left[(1-\alpha)R(i) + \alpha \max_{q \in \mathcal{C}_{i,a}} q^T V_2 \right] \\
&= \left(T^{\rm opt}_{\alpha} V_2\right)(i)
\end{aligned}$$

Proof (of Theorems 1 and 2). Let $\tilde{V}$ be the fixed point of $T^{\rm opt}_{\alpha,\mu}$. This means that for all $i \in S$,

$$\tilde{V}(i) = (1-\alpha)R(i) + \alpha \max_{q \in \mathcal{C}_{i,\mu(i)}} q^T \tilde{V} \ .$$

We wish to show that $\tilde{V} = V^{\rm opt}_{\alpha,\mu}$. Let q_i be the probability vector that achieves the maximum above. Construct an MDP $M_1 \in \mathcal{M}$ as follows. Set the transition probability vector $p_{i,\cdot}(\mu(i))$ to be q_i. For $a \neq \mu(i)$, choose $p_{i,\cdot}(a)$ to be any element of $\mathcal{C}_{i,a}$. It is clear that $\tilde{V}$ satisfies, for all $i \in S$,

$$\tilde{V}(i) = (1-\alpha)R(i) + \alpha \sum_j p_{i,j}(\mu(i)) \tilde{V}(j) \ ,$$

and therefore $\tilde{V} = V_{\alpha,\mu,M_1} \leq V^{\rm opt}_{\alpha,\mu}$. It remains to show that $\tilde{V} \geq V^{\rm opt}_{\alpha,\mu}$. For that, fix an arbitrary MDP $M \in \mathcal{M}$. Let V_0 be any initial vector. Using Lemma 11 and straightforward induction, we get

$$\forall k \geq 0, \ (T_{\alpha,\mu,M})^k V_0 \leq (T^{\rm opt}_{\alpha,\mu})^k V_0 \ .$$

Taking limits as $k \to \infty$, we get $V_{\alpha,\mu,M} \leq \tilde{V}$. Since $M \in \mathcal{M}$ was arbitrary, for any $i \in S$,

$$V^{\rm opt}_{\alpha,\mu}(i) = \sup_{M \in \mathcal{M}} V_{\alpha,\mu,M}(i) \leq \tilde{V}(i) \ .$$

Therefore, $\tilde{V} = V^{\rm opt}_{\alpha,\mu}$.

Now let $\tilde{V}$ be the fixed point of $T^{\rm opt}_{\alpha}$. This means that for all $i \in S$,

$$\tilde{V}(i) = \max_{a \in A} \left[(1-\alpha)R(i) + \alpha \max_{q \in \mathcal{C}_{i,a}} q^T \tilde{V} \right] \ .$$

We wish to show that $\tilde{V} = \mathbf{V}^{\rm opt}_{\alpha}$. Let $\mu_1(i)$ be any action that achieves the maximum above. Since $\tilde{V}$ satisfies, for all $i \in S$,

$$\tilde{V}(i) = (1-\alpha)R(i) + \alpha \max_{q \in \mathcal{C}_{i,\mu_1(i)}} q^T \tilde{V} \ ,$$

we have $\tilde{V} = V^{\text{opt}}_{\alpha,\mu_1} \leq \mathbf{V}^{\text{opt}}_{\alpha}$. It remains to show that $\tilde{V} \geq \mathbf{V}^{\text{opt}}_{\alpha}$. For that, fix an arbitrary policy μ. Let V_0 be any initial vector. Using Lemma 12 and straightforward induction, we get

$$\forall k \geq 0,\ (T^{\text{opt}}_{\alpha,\mu})^k V_0 \leq (T^{\text{opt}}_{\alpha})^k V_0\ .$$

Taking limits as $k \to \infty$, we get $V^{\text{opt}}_{\alpha,\mu} \leq \tilde{V}$. Since μ was arbitrary, for any $i \in S$,

$$\mathbf{V}^{\text{opt}}_{\alpha}(i) = \max_{\mu} V^{\text{opt}}_{\alpha,\mu}(i) \leq \tilde{V}(i)\ .$$

Therefore, $\tilde{V} = \mathbf{V}^{\text{opt}}_{\alpha}$. Moreover, this also proves the first part of Theorem 2 since

$$V^{\text{opt}}_{\alpha,\mu_1} = \tilde{V} = \mathbf{V}^{\text{opt}}_{\alpha}\ .$$

The claim that the fixed points of $T^{\text{pes}}_{\alpha,\mu}$ and T^{pes}_{α} are $V^{\text{pes}}_{\alpha,\mu}$ and $\mathbf{V}^{\text{pes}}_{\alpha}$ respectively, is proved by making a few obvious changes to the argument above. Further, as it turned out above, the argument additionally yields the proof of the second part of Theorem 2.

Proof of Theorem 3

We prove the existence of $\mathcal{M}_{\text{opt}}$ only. The existence of $\mathcal{M}_{\text{pes}}$ is proved in the same way. Note that in the proof presented in the previous subsection, given a policy μ, we explicitly constructed an MDP M_1 such that $V^{\text{opt}}_{\alpha,\mu} = V_{\alpha,\mu,M_1}$. Further, the transition probability vector $p_{i,\cdot}(\mu(i))$ of M_1 was a vector that achieved the maximum in

$$\max_{\mathcal{C}_{i,\mu(i)}} q^T V^{\text{opt}}_{\alpha,\mu}\ .$$

Recall that the set $\mathcal{C}_{i,\mu(i)}$ has the form

$$\{q\ :\ q^T\mathbf{1} = 1,\ \forall j \in S,\ l_j \leq q_j \leq u_j\}\ , \tag{10}$$

where $l_j = l(i,j,\mu(i))$, $u_j = u(i,j,\mu(i))$. Therefore, all that we require is the following lemma.

Lemma 13. *Given a set $\mathcal{C}$ of the form* (10)*, there exists a finite set $Q = Q(\mathcal{C})$ of cardinality no more than $|S|!$ with the following property. For any vector V, there exists $\tilde{q} \in Q$ such that*

$$\tilde{q}^T V = \max_{q \in \mathcal{C}} q^T V\ .$$

We can then set

$$\mathcal{M}_{\text{opt}} = \{\ \langle S, A, R, \{p_{i,j}(a)\}\rangle\ :\ \forall i, a,\ p_{i,\cdot}(a) \in Q(\mathcal{C}_{i,a})\ \}\ .$$

The cardinality of $\mathcal{M}_{\text{opt}}$ is at most $(|S||A|)|S|!$

Proof (of Lemma 13). A simple greedy algorithm (Algorithm 3) can be used to find a maximizing $\tilde{q}$. The set $\mathcal{C}$ is specified using upper and lower bounds, denoted by u_i and l_i respectively. The algorithm uses the following idea recursively. Suppose i^* is the index of a largest component of V. It is clear that we should set $\tilde{q}(i^*)$ as large as possible. The value of $\tilde{q}(i^*)$ has to be less than u_i. Moreover, it has to be less than $1 - \sum_{i \neq i^*} l_i$. Otherwise, the remaining lower bound constraints cannot be met. So, we set $\tilde{q}(i^*)$ to be the minimum of these two quantities.

Note that the output depends only on the sorted order of the components of V. Hence, there are only $|S|!$ choices for $\tilde{q}$.

Algorithm 3. A greedy algorithm to maximize $q^T V$ over $\mathcal{C}$.

INPUTS The vector V and the set $\mathcal{C}$. The latter is specified by bounds $\{l_i\}_{i \in S}$ and $\{u_i\}_{i \in S}$ that satisfy $\forall i, 0 \leq l_i \leq u_i$ and $\sum_i l_i \leq 1 \leq \sum_i u_i$.
OUTPUT A maximizing vector $\tilde{q} \in \mathcal{C}$.

$indices \leftarrow$ **order**(V) ▷ **order**(V) gives the indices of the largest to smallest elements of V

$massLeft \leftarrow 1$
$indicesLeft \leftarrow S$
for all $i \in indices$ **do**
  $elem \leftarrow V(i)$
  $lowerBoundSum \leftarrow \sum_{j \in indicesLeft, j \neq i} l_j$
  $\tilde{q}(i) \leftarrow \min(u_i, massLeft - lowerBoundSum)$
  $massLeft \leftarrow massLeft - \tilde{q}(i)$
  $indicesLeft \leftarrow indicesLeft - \{i\}$
end for
return $\tilde{q}$