Bounded Parameter Markov Decision Processes

Robert Givan, Sonia Leach, Thomas Dean

Department of Computer Science
Brown University
Providence, Rhode Island 02912

CS-97-05
May 1997

Bounded Parameter Markov Decision Processes

Robert Givan Sonia Leach Thomas Dean

Department of Computer Science, Brown University
115 Waterman Street, Providence, RI 02912, USA
Phone: (401) 863-7600 Fax: (401) 863-7657
Email: {rlg,sml,tld}@cs.brown.edu

Abstract

In this paper, we introduce the notion of a *bounded parameter Markov decision process* as a generalization of the traditional *exact* MDP. A bounded parameter MDP is a set of exact MDPs specified by giving upper and lower bounds on transition probabilities and rewards (all the MDPs in the set share the same state and action space). Bounded parameter MDPs can be used to represent variation or uncertainty concerning the parameters of sequential decision problems. Bounded parameter MDPs can also be used in aggregation schemes to represent the variation in the transition probabilities for different base states aggregated together in the same aggregate state.

We introduce *interval value functions* as a natural extension of traditional value functions. An interval value function assigns a closed real interval to each state, representing the assertion that the value of that state falls within that interval. An interval value function can be used to bound the performance of a policy over the set of exact MDPs associated with a given bounded parameter MDP. We describe an iterative dynamic programming algorithm called *interval policy evaluation* which computes an interval value function for a given bounded parameter MDP and specified policy. Interval policy evaluation on a policy π computes the most restrictive interval value function that is sound, *i.e.*, that bounds the value function for π in every exact MDP in the set defined by the bounded parameter MDP. A simple modification of interval policy evaluation results in a variant of value iteration [Bellman, 1957] that we call *interval value iteration* which computes a policy for an bounded parameter MDP that is optimal in a well-defined sense.

Keywords: Decision-theoretic planning, Planning under uncertainty, Approximate planning, Markov decision processes.

1 Introduction

The theory of Markov decision processes (MDPs) provides the semantic foundations for a wide range of problems involving planning under uncertainty [Boutilier *et al.*, 1995a, Littman, 1997]. In this paper, we introduce a generalization of Markov decision processes called *bounded parameter Markov decision processes* (BMDPs) that allows us to model uncertainty in the parameters that comprise an MDP. Instead of encoding a parameter such as the probability of making a transition from one state to another as a single number, we specify a range of possible values for the parameter as a closed interval of the real numbers.

A BMDP can be thought of as a family of traditional (exact) MDPs, *i.e.,* the set of all MDPs whose parameters fall within the specified ranges. From this perspective, we may have no justification for committing to a particular MDP in this family, and wish to analyze the consequences of this lack of commitment. Another interpretation for a BMDP is that the states of the BMDP actually represent sets (aggregates) of more primitive states that we choose to group together. The intervals here represent the ranges of the parameters over the primitive states belonging to the aggregates.

In a related paper, we have shown how BMDPs can be used as part of a strategy for efficiently approximating the solution of MDPs with very large state spaces and dynamics compactly encoded in a factored (or implicit) representation [Dean *et al.*, 1997]. In this paper, we focus exclusively on BMDPs, on the BMDP analog of value functions, called *interval value functions*, and on policy selection for a BMDP. We provide BMDP analogs of the standard (exact) MDP algorithms for computing the value function for a fixed policy (plan) and (more generally) for computing the optimal value function over all policies, called *interval policy evaluation* and *interval value iteration* (IVI) respectively. We define the desired output values for these algorithms and prove that the algorithms converge to these desired values. Finally, we consider two different notions of optimal policy for an BMDP, and show how IVI can be applied to extract the optimal policy for each notion. The first notion of optimality states that the desired policy must perform better than any other under the assumption that an adversary selects the model parameters. The second notion requires the best possible performance when a friendly choice of model parameters is assumed.

2 Exact Markov Decision Processes

An (exact) Markov decision process M is a four tuple $M = (\mathcal{Q}, \mathcal{A}, F, R)$ where $\mathcal{Q}$ is a set of states, $\mathcal{A}$ is a set of actions, R is a reward function that

maps each state to a real value $R(q)$,[1] and F is a state-transition distribution so that for $\alpha \in \mathcal{A}$ and $p, q \in \mathcal{Q}$,

$$F_{pq}(\alpha) = \Pr(X_{t+1} = q | X_t = p, U_t = \alpha)$$

where X_t and U_t are random variables denoting, respectively, the state and action at time t. In cases where it is necessary to make explicit from which MDP a transition function originates, let F^M denote the transition function of the MDP M.

A *policy* is a mapping from states to actions, $\pi : \mathcal{Q} \to \mathcal{A}$. The set of all policies is denoted Π. An MDP M together with a fixed policy $\pi \in \Pi$ determines a Markov chain such that the probability of making a transition from p to q is defined by $F_{pq}(\pi(p))$. The *expected value function* (or simply the *value function*) associated with such a Markov chain is denoted $V_{M,\pi}$. The value function maps each state to its *expected discounted cumulative reward* defined by

$$V_{M,\pi}(p) = R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\pi(p)) V_{M,\pi}(q)$$

where $0 \leq \gamma < 1$ is called the *discount rate*.[2] In most contexts, the relevant MDP is clear and we abbreviate $V_{M,\pi}$ as V_π.

The optimal value function V^*_M (or simply V^* where the relevant MDP is clear) is defined as follows.

$$V^*(p) = \max_{\alpha \in \mathcal{A}} \left(R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\alpha) V^*(q) \right)$$

The value function V^* is greater than or equal to any value function V_π in the partial order $\geq_{\text{dom}}$ defined as follows: $V_1 \geq_{\text{dom}} V_2$ if and only if for all states q, $V_1(q) \geq V_2(q)$.

An optimal policy is any policy π^* for which $V^* = V_{\pi^*}$. Every MDP has at least one optimal policy, and the set of optimal policies can be found by replacing the max in the definition of V^* with arg max.

3 Bounded Parameter Markov Decision Processes

A *bounded parameter MDP* (BMDP) is a four tuple $\mathcal{M} = (\mathcal{Q}, \mathcal{A}, \hat{F}, \hat{R})$ where $\mathcal{Q}$ and $\mathcal{A}$ are defined as for MDPs, and $\hat{F}$ and $\hat{R}$ are analogous to the MDP

[1] The techniques and results in this paper easily generalize to more general reward functions. We adopt a less general formulation to simplify the presentation.

[2] In this paper, we focus on expected discounted cumulative reward as a performance criterion, but other criteria, *e.g.*, total or average reward [Puterman, 1994], are also applicable to bounded parameter MDPs.

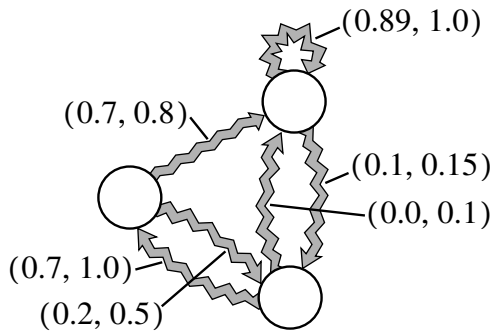

Figure 1: The state-transition diagram for a simple bounded parameter Markov decision process with three states and a single action. The arcs indicate possible transitions and are labeled by their lower and upper bounds.

F and R but yield closed real intervals instead of real values. That is, for any action α and states p, q, $\hat{R}(p)$ and $\hat{F}_{p,q}(\alpha)$ are both closed real intervals of the form $[l, u]$ for l and u real numbers with $l \leq u$.[3] To ensure that $\hat{F}$ admits only well-formed transition functions, we require that $l \geq 0$ and $u \leq 1$. Also, for any action α and state p, the sum of the lower bounds of $\hat{F}_{pq}(\alpha)$ over all states q must be less than or equal to 1 while the upper bounds must sum to a value greater than or equal to 1. Figure 1 depicts the state-transition diagram for a simple BMDP with three states and one action.

A BMDP $\mathcal{M} = (\mathcal{Q}, \mathcal{A}, \hat{F}, \hat{R})$ defines a set of exact MDPs which, by abuse of notation, we also call $\mathcal{M}$. For exact MDP $M = (\mathcal{Q}', \mathcal{A}', F', R')$, we have $M \in \mathcal{M}$ if $\mathcal{Q} = \mathcal{Q}'$, $\mathcal{A} = \mathcal{A}'$, and for any action α and states p, q, $R'(p)$ is in the interval $\hat{R}(p)$ and $F'_{p,q}(\alpha)$ is in the interval $\hat{F}_{p,q}(\alpha)$. We rely on context to distinguish between the tuple view of $\mathcal{M}$ and the exact MDP set view of $\mathcal{M}$. In the definitions in this section, the BMDP $\mathcal{M}$ is implicit.

An *interval value function* $\hat{V}$ is a mapping from states to closed real intervals. We generally use such functions to indicate that the given state's value falls within the selected interval. Interval value functions can be specified for both exact and BMDPs. As in the case of (exact) value functions, interval value functions are specified with respect to a fixed policy. Note that in the case of BMDPs a state can have a range of values depending on how the transition and reward parameters are instantiated, hence the need for an interval value function.

For each of the interval valued functions $\hat{F}$, $\hat{R}$, $\hat{V}$ we define two real valued functions which take the same arguments and give the upper and lower interval bounds, denoted $\overline{F}$, $\overline{R}$, $\overline{V}$, and $\underline{F}$, $\underline{R}$, $\underline{V}$, respectively. So, for example, at any state q we have $\hat{V}(q) = [\underline{V}(q), \overline{V}(q)]$.

[3] To simplify the remainder of the paper, we assume that the reward bounds are always tight, *i.e.*, that for all $q \in \mathcal{Q}$, for some real l, $\hat{R}(q) = (l, l)$, and we refer to l as $R(q)$. The generalization to nontrivial bounds on rewards is straightforward.

Definition 1 *For any policy π and state q, we define the interval value $\hat{V}_\pi(q)$ of π at q to be the interval*

$$\left[\min_{M\in\mathcal{M}} V_{M,\pi}(q),\ \max_{M\in\mathcal{M}} V_{M,\pi}(q)\right]$$

In Section 5 we will give an iterative algorithm which we have proven to converge to $\hat{V}_\pi$. In preparation for that discussion, we will show in Theorem 1 that there is at least one specific MDP in $\mathcal{M}$ which simultaneously achieves $\overline{V}_\pi(q)$ for all states q (and likewise a specific MDP achieving $\underline{V}_\pi(q)$ for all q). We formally define such an MDP below.

Definition 2 *For any policy π, an MDP in $\mathcal{M}$ is* π-maximizing *if it is a possible value of* $\arg\max_{M\in\mathcal{M}} V_{M,\pi}$ *and it is* π-minimizing *if it is in* $\arg\min_{M\in\mathcal{M}} V_{M,\pi}$.

The proof for Theorem 1 builds on the fact that for a particular policy π, if there are MDPs M_1 and M_2 that minimize (maximize) the value function for particular states p_1 and p_2, respectively, then we can compose a new MDP which minimizes (maximizes) the value function for both p_1 and p_2 given π. By repeatedly composing MDPs, we can create a π-minimizing (π-maximizing) MDP. The following definition and two lemmas formalize this argument.

Definition 3 *For any two MDPs M_1, M_2, and policy $\pi \in \Pi$, we define the composition operator $\bigcirc_M$ for MDPs as $M_3 = M_1 \bigcirc_M M_2$ if for all states $p, q \in \mathcal{Q}$,*

$$F_{pq}^{M_3}(\pi(p)) = \begin{cases} F_{pq}^{M_1}(\pi(p)) & \text{if } V_{M_1,\pi}(p) \geq V_{M_2,\pi}(p) \\ F_{pq}^{M_2}(\pi(p)) & \text{otherwise} \end{cases}$$

Lemma 1 *If $M_3 = M_1 \bigcirc_M M_2$, then*

$$V_{M_3,\pi} \geq_{\text{dom}} V_{M_1,\pi} \quad \text{and} \quad V_{M_3,\pi} \geq_{\text{dom}} V_{M_2,\pi}.$$

Proof: We prove the theorem by constructing a value function v such that $v \geq_{\text{dom}} V_{M_1,\pi}$ and $v \geq_{\text{dom}} V_{M_2,\pi}$. We then show that $v \leq_{\text{dom}} V_{M_3,\pi}$ using We construct v as follows. Let

$$v(p) = \max\{V_{M_1,\pi}(p), V_{M_2,\pi}(p)\} \tag{1}$$

for all $p \in \mathcal{Q}$. Note that this implies $v \geq_{\text{dom}} V_{M_1,\pi}$ and $v \geq_{\text{dom}} V_{M_2,\pi}$. We now show using (8) that $v \leq_{\text{dom}} V_{M_3,\pi}$.

We divide the set of states $\mathcal{Q}$ into two subsets. Let $\mathcal{Q}_1$ be the set of states $\{q\}$ whose value estimate $v(q) = V_{M_1,\pi}(q)$ and let $\mathcal{Q}_2$ be the set of states $\{q\}$

whose value estimate $v(q) = V_{M_2,\pi}(q)$ as determined by (1). For MDP M_1, we have $V_{M_1,\pi}(q) = v(q)$ for all $q \in \mathcal{Q}_1$ and $V_{M_1,\pi}(q) \leq v(q)$ for all $q \in \mathcal{Q}_2$.

Suppose without loss of generality that we choose a particular state $p \in \mathcal{Q}_1$. Then $v(p) = V_{M_1,\pi}(p)$ and by the definition of $\bigcirc_M$, we know $F_{pq}^{M}(\pi(p)) = F_{pq}^{M_1}(\pi(p))$ for all $q \in \mathcal{Q}$. Furthermore, for the value functions v and $V_{M_1,\pi}$, we have for the states in $\mathcal{Q}_1$,

$$\sum_{q \in \mathcal{Q}_1} \left(F_{pq}^{M_1}(\pi(p)) V_{M_1,\pi}(q) \right) = \sum_{q \in \mathcal{Q}_1} \left(F_{pq}^{M_1}(\pi(p)) v(q) \right)$$

and for the states in $\mathcal{Q}_2$,

$$\sum_{q \in \mathcal{Q}_2} \left(F_{pq}^{M_1}(\pi(p)) V_{M_1,\pi}(q) \right) \leq \sum_{q \in \mathcal{Q}_2} \left(F_{pq}^{M_1}(\pi(p)) v(q) \right).$$

Using these facts, we can write

$$\begin{aligned} v(p) &= V_{M_1,\pi}(p) \\ &= R(p) + \gamma \sum_{q \in \mathcal{Q}} \left(F_{pq}^{M_1}(\pi(p)) V_{M_1,\pi}(q) \right) \\ &\leq R(p) + \gamma \sum_{q \in \mathcal{Q}} \left(F_{pq}^{M_1}(\pi(p)) v(q) \right) \\ &= R(p) + \gamma \sum_{q \in \mathcal{Q}} \left(F_{pq}^{M_3}(\pi(p)) v(q) \right) \\ &= V I_{M_3,\pi(p)}(v)(p). \end{aligned}$$

It follows from Theorem 8 that $v(p) \leq V_{M_3,\pi}(p)$. Repeating this argument in the case that $p \in \mathcal{Q}_2$ implies $v \leq_{\text{dom}} V_{M_3,\pi}$. $\square$

We now use the MDP composition operator to show that we can apply it repeatedly to a subset K of states to create a π-minimizing (π-maximizing) MDP over the subset K.

Lemma 2 *For any policy π and any $K \subseteq \mathcal{Q}$, there exist π-maximizing and π-minimizing MDPs in $\mathcal{M}$ on the subset K.*

Proof: We prove the lemma for the π-maximizing MDP by induction on the size k of the subset K. The proof for the π-minimizing MDP is analogous.

Base Case: ($k = 1$) THIS ONE STILL NEEDS TO BE PROVED, EITHER BY SHOWING THE QUADRATIC PROGRAM TO HAVE A SOLUTION OR USING THE UPPER SEMICONTINUOUS FUNCTION ARGUMENT TO SHOW THE MAX IS ACHIEVED.

Inductive Case: ($k > 1$) Assume the theorem holds for $n < k$. We divide the subset K of states into two disjoint, non-empty subsets K_1 and K_2, where $K_1 \cup K_2 = K$. Then there exist MDPs M_1 and M_2 for which

the inductive hypothesis holds for K_1 and K_2, respectively. In particular, for M_1,

$$V_{M_1,\pi}(p) \geq V_{M,\pi}(p)$$

where $p \in K_1$ and $M \in \mathcal{M}$. Similarly for $V_{M_2,\pi}$ on the states $p \in K_2$. Let M_3 be the MDP given as $M_3 = M_1 \bigcirc_M M_2$. From Lemma 1, we know that $V_{M_1,\pi} \leq_{\text{dom}} V_{M_3,\pi}$ and $V_{M_2,\pi} \leq_{\text{dom}} V_{M_3,\pi}$ which implies that $V_{M,\pi}(p) \leq V_{M_3,\pi}(p)$ for all states $p \in K$ and $M \in \mathcal{M}$. □

Theorem 1 *For any policy π, there exist π-maximizing and π-minimizing MDPs in $\mathcal{M}$.*

Proof: The proof follows immediately from Lemma 2 by letting $K = \mathcal{Q}$.

This theorem implies that $\underline{V}_\pi$ is equivalent to $\min_{M \in \mathcal{M}} V_{M,\pi}$ where the minimization is done relative to $\geq_{\text{dom}}$, and likewise for $\overline{V}$ using max. We give an algorithm in Section 5 which converges to $\underline{V}_\pi$ by also converging to a π-minimizing MDP in $\mathcal{M}$ (likewise for $\overline{V}_\pi$).

We now consider how to define an optimal value function for an BMDP. Consider the expression $\max_{\pi \in \Pi} \hat{V}_\pi$. This expression is ill-formed because we have not defined how to rank the interval value functions $\hat{V}_\pi$ in order to select a maximum. We focus here on two different ways to order these value functions, yielding two notions of optimal value function and optimal policy. Other orderings may also yield interesting results.

First, we define two different orderings on closed real intervals:

$$[l_1, u_1] \leq_{\text{pes}} [l_2, u_2] \iff \begin{cases} l_1 < l_2, & \text{or} \\ l_1 = l_2 & \text{and } u_1 \leq u_2 \end{cases}$$

$$[l_1, u_1] \leq_{\text{opt}} [l_2, u_2] \iff \begin{cases} u_1 < u_2, & \text{or} \\ u_1 = u_2 & \text{and } l_1 \leq l_2 \end{cases}$$

We extend these orderings to partially order interval value functions by relating two value functions $\hat{V}_1 \leq \hat{V}_2$ only when $\hat{V}_1(q) \leq \hat{V}_2(q)$ for every state q. We can now use either of these orderings to compute $\max_{\pi \in \Pi} \hat{V}_\pi$, yielding two definitions of optimal value function and optimal policy. However, since the orderings are partial (on value functions), we must still prove that the set of policies contains a policy which achieves the desired maximum under each ordering (*i.e.,* a policy whose interval value function is ordered above that of every other policy). We formally define such policies below.

Definition 4 *The* optimistic optimal value function $\hat{V}_{\text{opt}}$ *and the* pessimistic optimal value function $\hat{V}_{\text{pes}}$ *are given by:*

$$\hat{V}_{\text{opt}} = \max_{\pi \in \Pi} \hat{V}_\pi \quad \textit{using } \leq_{\text{opt}} \textit{ to order interval value functions}$$
$$\hat{V}_{\text{pes}} = \max_{\pi \in \Pi} \hat{V}_\pi \quad \textit{using } \leq_{\text{pes}} \textit{ to order interval value functions}$$

We say that any policy π whose interval value function $\hat{V}_\pi$ is $\geq_{\text{opt}}$ ($\geq_{\text{pes}}$) the value functions $\hat{V}_{\pi'}$ of all other policies π' is optimistically (pessimistically) optimal.

The proof for Theorem 2 is similar to that of Theorem 1. Analogous to the composition operator for MDPs, we define a composition operator on policies. Repeatedly composing policies, as we did for MDPs, we can create a policy that is optimistically (pessimistically) optimal. The following definition and ensuing pair of lemmas provide the necessary material for the proof of Theorem 2.

Definition 5 *For any two policies $\pi_1, \pi_2 \in \Pi$ and MDP M, we define the composition operator $\bigcirc_\pi$ on policies as $\pi_3 = \pi_1 \bigcirc_\pi \pi_2$ if for all states $p \in \mathcal{Q}$,:*

$$\pi_3(p) = \begin{cases} \pi_1(p) & \text{if } V_{M,\pi_1}(p) \geq V_{M,\pi_2}(p) \\ \pi_2(p) & \text{otherwise} \end{cases}$$

Lemma 3 *If $\pi_3 = \pi_1 \bigcirc_\pi \pi_2$, then $V_{M,\pi_3} \geq_{\text{dom}} V_{M,\pi_1}$ and $V_{M,\pi_3} \geq_{\text{dom}} V_{M,\pi_2}$.*

Proof: We prove the theorem by constructing a value function v such that $v \geq_{\text{dom}} V_{M,\pi_1}$ and $v \geq_{\text{dom}} V_{M,\pi_2}$. We then show that $v \leq_{\text{dom}} V_{M,\pi_3}$ using Theorem 8. We construct v as follows. Let

$$v(p) = \max\{V_{M,\pi_1}(p), V_{M,\pi_2}(p)\} \tag{2}$$

for all $p \in \mathcal{Q}$. Note that this implies $v \geq_{\text{dom}} V_{M,\pi_1}$ and $v \geq_{\text{dom}} V_{M,\pi_2}$. We now show using (8) that $v \leq_{\text{dom}} V_{M,\pi_3}$.

We divide the set of states $\mathcal{Q}$ into two subsets. Let $\mathcal{Q}_1$ be the set of states $\{q\}$ whose value estimate $v(q) = V_{M,\pi_1}(q)$ and let $\mathcal{Q}_2$ be the set of states $\{q\}$ whose value estimate $v(q) = V_{M,\pi_2}(q)$ as determined by (2). For policy π_1, we have $V_{M,\pi_1}(q) = v(q)$ all states $q \in \mathcal{Q}_1$ and $V_{M,\pi_1}(q) \leq v(q)$ for all states $q \in \mathcal{Q}_2$.

Suppose without loss of generality that we choose a particular state $p \in \mathcal{Q}_1$. Then $v(p) = V_{M,\pi_1}(p)$ and by the definition of $\bigcirc_\pi$, $\pi_3(p) = \pi_1(p)$. Using these facts, we can write

$$\begin{aligned} v(p) &= V_{M,\pi_1}(p) \\ &= R(p) + \gamma \sum_{q \in \mathcal{Q}} (F_{pq}(\pi_1(p)) V_{M,\pi_1}(q)) \\ &\leq R(p) + \gamma \sum_{q \in \mathcal{Q}} (F_{pq}(\pi_1(p)) v(q)) \\ &= R(p) + \gamma \sum_{q \in \mathcal{Q}} (F_{pq}(\pi_3(p)) v(q)) \\ &= VI_{M,\pi_3}(v)(p). \end{aligned}$$

It follows from Theorem 8 that $v(p) \leq V_{M,\pi_3}(p)$. Repeating this argument in the case that $p \in \mathcal{Q}_2$ implies $v \leq_{\text{dom}} V_{M,\pi_3}$. □

Lemma 4 *For any subset K of $\mathcal{Q}$, there exists a policy π' such that*

$$\hat{V}_{\pi}(p) \leq_{\text{opt}} (\leq_{\text{pes}}) \hat{V}_{\pi'}(p)$$

for all states $p \in K$ and $\pi \in \Pi$.

Proof: We prove the lemma by induction on the size k of the subset K.

Base Case: $(k = 1)$ Since $\mathcal{A}$ and $\mathcal{Q}$ are finite sets, Π is a finite set. As a consequence of Theorem 1, we know we can obtain a finite set of interval value functions $\{\hat{V}_{\pi} : \pi \in \Pi\}$. Any finite set of intervals has a maximum under either ordering $\leq_{\text{pes}}$ or $\leq_{\text{opt}}$.

Inductive Case: $(k > 1)$ Assume the theorem holds for $n < k$. We divide the subset K of states into two disjoint, non-empty subsets K_1 and K_2, where $K_1 \cup K_2 = K$. Then there exist policies π_1 and π_2 for which the inductive hypothesis holds for K_1 and K_2, respectively. In particular, for π_1,

$$\hat{V}_{\pi}(p) \leq_{\text{opt}} (\leq_{\text{pes}}) \hat{V}_{\pi_1}(p)$$

where $p \in K_1$ and $\pi \in \Pi$. Similarly for $\hat{V}_{\pi_2}$ on the states $p \in K_2$. Let π_3 be the policy given as $\pi_3 = \pi_1 \bigcirc_{\pi} \pi_2$. We show that $\hat{V}_{\pi_1} \leq_{\text{opt}} (\leq_{\text{pes}}) \hat{V}_{\pi_3}$ and $\hat{V}_{\pi_2} \leq_{\text{opt}} (\leq_{\text{pes}}) \hat{V}_{\pi_3}$, which implies $\hat{V}_{\pi}(p) \leq_{\text{opt}} (\leq_{\text{pes}}) \hat{V}_{\pi_3}(p)$ for all states $p \in K$ and $\pi \in \Pi$.

We argue the case for $\hat{V}_{\pi_1}$; the case for $\hat{V}_{\pi_2}$ is analogous. For each $M \in \mathcal{M}$,

$$V_{M,\pi_1} \leq_{\text{dom}} V_{M,\pi_3}$$

by Lemma 3. It follows for all states $p \in \mathcal{Q}$ that,

$$\min_{M \in \mathcal{M}} V_{M,\pi_1}(p) \leq \min_{M \in \mathcal{M}} V_{M,\pi_3}(p)$$

and

$$\max_{M \in \mathcal{M}} V_{M,\pi_1}(p) \leq \max_{M \in \mathcal{M}} V_{M,\pi_3}(p).$$

Note that, as a consequence, the intervals $\hat{V}_{\pi_1}(p)$ and $\hat{V}_{\pi_3}(p)$ may overlap, but one can not contain the other. Therefore, according to either ordering $\leq_{\text{opt}}$ or $\leq_{\text{pes}}$,

$$\hat{V}_{\pi_1} \leq_{\text{opt}} (\leq_{\text{pes}}) \hat{V}_{\pi_3}.$$

□

Theorem 2 *There exists at least one optimistically (pessimistically) optimal policy, and therefore the definition of $\hat{V}_{\text{opt}}$ ($\hat{V}_{\text{pes}}$) is well-formed.*

Proof: The proof follows immediately from Lemma 4 by letting $K = \mathcal{Q}$. □

The above two notions of optimal value can be understood in terms of a game in which we choose a policy π and then a second player chooses in which MDP M in $\mathcal{M}$ to evaluate the policy. The goal is to get the highest[4] resulting value function $V_{M,\pi}$. The optimistic optimal value function's upper bounds $\overline{V}_{\text{opt}}$ represent the best value function we can obtain in this game if we assume the second player is cooperating with us. The pessimistic optimal value function's lower bounds $\underline{V}_{\text{pes}}$ represent the best we can do if we assume the second player is our adversary, trying to minimize the resulting value function.

In the next section, we describe well-known iterative algorithms for computing the optimal value function V^*, and then in Section 5 we will describe similar iterative algorithms which compute $\hat{V}_{\text{opt}}$ ($\hat{V}_{\text{pes}}$).

4 Estimating Traditional Value Functions

In this section, we review the basics concerning dynamic programming methods for computing value functions for fixed and optimal policies in traditional MDPs. In the next section, we describe novel algorithms for computing the interval analogs of these value functions for bounded parameter MDPs.

We present results from the theory of exact MDPs which rely on the concept of normed linear spaces. We define operators, VI_π and VI, on the space of value functions. We then use the Banach fixed-point theorem (Theorem 3) to show that iterating these operators converges to unique fixed-points, V_π and V^* respectively (Theorems 5 and 6).

Let $\mathcal{V}$ denote the set of value functions on $\mathcal{Q}$. For each $v \in \mathcal{V}$, define the (sup) *norm* of v by

$$\|v\| = \max_{q \in \mathcal{Q}} |v(q)|.$$

We use the term *convergence* to mean convergence in the norm sense. The space $\mathcal{V}$ together with $\|\cdot\|$ constitute a complete normed linear space, or *Banach Space*. If U is a Banach space, then an operator $T : U \to U$ is a *contraction mapping* if there exists a λ, $0 \leq \lambda < 1$ such that $\|Tv - Tu\| \leq \lambda\|v - u\|$ for all u and v in U.

Define $VI : \mathcal{V} \to \mathcal{V}$ and for each $\pi \in \Pi$, $VI_\pi : \mathcal{V} \to \mathcal{V}$ on each $p \in \mathcal{Q}$ by

$$\begin{aligned} VI(v)(p) &= \max_{\alpha \in \mathcal{A}} \left(R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\alpha) v(q) \right) \\ VI_\pi(v)(p) &= R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\pi(p)) v(q). \end{aligned}$$

[4] Value functions are ranked by $\geq_{\text{dom}}$.

In cases where it is necessary for the definitions of VI and VI_π to make explicit from which MDP the transition function originates, we write $VI_{M,\pi}$ and VI_M to denote the operators VI_π and VI as defined previously, except that the transition function F is F^M. More generally, we write $VI_{M,\alpha} : \mathcal{V} \to \mathcal{V}$ to denote the operator defined on each $p \in \mathcal{Q}$ as

$$VI_{M,\alpha}(v)(p) = R(p) + \gamma \sum_{q \in \mathcal{Q}} F^M_{pq}(\alpha) v(q)$$

Using these operators, we can rewrite the expression for V^* and V_π as

$$V^*(p) = VI(V^*)(p) \quad \text{and} \quad V_\pi(p) = VI_\pi(V_\pi)(p)$$

for all states $p \in \mathcal{Q}$. This implies that V^* and V_π are fixed points of VI and VI_π, respectively. The following four theorems show that for each operator, iterating the operator on an initial value estimate converges to these fixed points.

Theorem 3 *For any Banach space U and contraction mapping $T : U \to U$, there exists a unique v^* in U such that $Tv^* = v^*$; and for arbitrary v^0 in U, the sequence $\{v^n\}$ defined by $v^n = Tv^{n-1} = T^n v^0$ converges to v^*.*

Theorem 4 *VI and VI_π are contraction mappings.*

Theorem 3 and Theorem 4 together prove the following fundamental results in the theory of MDPs.

Theorem 5 *There exists a unique $v^* \in \mathcal{V}$ satisfying $v^* = VI(v^*)$; furthermore, $v^* = V^*$. Similarly, V_π is the unique fixed-point of VI_π.*

Theorem 6 *For arbitrary $v^0 \in \mathcal{V}$, the sequence $\{v^n\}$ defined by $v^n = VI(v^{n-1})$ $= VI^n(v^0)$ converges to V^*. Similarly, iterating VI_π converges to V_π.*

An important consequence of Theorem 6 is that it provides an algorithm for finding V^* and V_π. In particular, to find V^*, we can start from an arbitrary initial value function v^0 in $\mathcal{V}$, and repeatedly apply the operator VI to obtain the sequence $\{v^n\}$. This algorithm is referred to as *value iteration*. Theorem 6 guarantees the convergence of value iteration to the optimal value function. Similarly, we can specify an algorithm called *policy evaluation* which finds V_π by repeatedly apply VI_π starting with an initial $v^0 \in \mathcal{V}$.

The following theorem from [Littman *et al.*, 1995] states the convergence rate of value iteration and policy evaluation.

Theorem 7 *For fixed γ, value iteration and policy evaluation converge to the optimal value function in a number of steps polynomial in the number of states, the number of actions, and the number of bits used to represent the MDP parameters.*

Another important theorem that will be used in a number of later proofs shows how the result of applying the operators to a value function relates to the value of the operators' fixed-points.

Theorem 8 *Let $u \in \mathcal{V}$, $\pi \in \Pi$ and M be an MDP. For all states $p \in \mathcal{Q}$, if $u(p) \leq (\geq) VI_{M,\pi}(u)(p)$ then $u(p) \leq (\geq) V_{M,\pi}(p)$.*

5 Estimating Interval Value Functions

In this section, we describe dynamic programming algorithms which operate on bounded parameter MDPs. We first define the interval equivalent of policy evaluation $I\hat{V}I_\pi$ which computes $\hat{V}_\pi$, and then define the variants $I\hat{V}I_{opt}$ and $I\hat{V}I_{pes}$ which compute the optimistic and pessimistic optimal value functions.

5.1 Interval Policy Evaluation

In direct analogy to the definition of VI_π in Section 4, we define a function $I\hat{V}I_\pi$ (for *interval value iteration*) which maps interval value functions to other interval value functions. We have proven that iterating $I\hat{V}I_\pi$ on any initial interval value function produces a sequence of interval value functions which converges[5] to $\hat{V}_\pi$.

$I\hat{V}I_\pi(\hat{V})$ is an interval value function, defined for each state p as follows:

$$I\hat{V}I_\pi(\hat{V})(p) \quad = \quad \left[\min_{M \in \mathcal{M}} VI_{M,\pi(p)}(\underline{V})(p) \;\; \max_{M \in \mathcal{M}} VI_{M,\pi(p)}(\overline{V})(p)\right].$$

Associated with $I\hat{V}I_\pi$ are the functions $\underline{IVI}_\pi$ and $\overline{IVI}_\pi$ which are mappings from interval value functions to value functions. $\underline{IVI}_\pi(\hat{V})$ is the value function derived by extracting the lower bound of $I\hat{V}I_\pi(\hat{V})(p)$ for all states p. Similarly for $\overline{IVI}_\pi(\hat{V})$ on the upper bounds.

The algorithm to compute $I\hat{V}I|\pi$ is very similar to the standard MDP computation of VI, except that we must now be able to select an MDP M from the family $\mathcal{M}$ which minimizes (maximizes) the value attained. We select such an MDP by selecting a function F within the bounds specified by $\hat{F}$ to minimize (maximize) the value—each possible way of selecting F corresponds to one MDP in $\mathcal{M}$. We can select the values of $\hat{F}_{pq}(\alpha)$ independently for each α and p, but the values selected for different states q (for fixed α and p) interact: they must sum up to one. We now show how to determine, for

[5] We conjecture that this convergence is complete in time polynomial in the input size, following the analogous theorem for standard MDP policy evaluation[Littman *et al.*, 1995]. This result should be proven by publication time.

fixed α and p, the value of $\hat{F}_{pq}(\alpha)$ for each state q so as to minimize (maximize) the expression $\sum_{q \in \mathcal{Q}} (F_{pq}(\alpha)V(q))$. This step constitutes the heart of the IVI algorithm and the only significant way the algorithm differs from standard value iteration.

The idea is to sort the possible destination states q into increasing (decreasing) order according to their $\underline{V}$ ($\overline{V}$) value, and then choose the transition probabilities within the intervals specified by $\hat{F}$ so as to send as much probability mass to the states early in the ordering. Let $q_1, q_2, \ldots, q_k$ be such an ordering of $\mathcal{Q}$—so that, in the minimizing case, for all i and j if $1 \leq i \leq j \leq k$ then $\underline{V}(q_i) \leq \underline{V}(q_j)$ (increasing order).

Let r be the index $1 \leq r \leq k$ which maximizes the following expression without letting it exceed 1:

$$\sum_{i=1}^{r-1} \overline{F}_{p,q_i}(\alpha) + \sum_{i=r}^{k} \underline{F}_{p,q_i}(\alpha)$$

r is the index into the sequence q_i such that below index r we can assign the upper bound, and above index r we can assign the lower bound, with the rest of the probability mass from p under α being assigned to q_r. Formally, we choose $F_{pq}(\alpha)$ for all $q \in \mathcal{Q}$ as follows:

$$\begin{aligned} F_{pq_j}(\alpha) &= \begin{cases} \overline{F}_{p,q_i}(\alpha) & \text{if} \quad j < r \\ \underline{F}_{p,q_i}(\alpha) & \text{if} \quad j > r \end{cases} \\ F_{pq_r}(\alpha) &= 1 - \sum_{i=1, i \neq r}^{i=k} F_{pq_i}(\alpha) \end{aligned}$$

Figure 2 illustrates the basic iterative step in the above algorithm, for the maximizing case. The states q_i are ordered according to the value estimates in $\overline{V}$. The transitions from a state p to states q_i are defined by the function F such that each transition is equal to its lower bound plus some fraction of the leftover probability mass.

Techniques similar to those in Section 4 can be used to prove that iterating $\underline{IVI}_\pi$ ($\overline{IVI}_\pi$) converges to $\underline{V}_\pi$ ($\overline{V}_\pi$). The key theorems, stated below, assert first that $\underline{IVI}_\pi$ is a contraction mapping, and second that $\underline{V}_\pi$ is a fixed-point of $\underline{IVI}_\pi$, and are easily proven.

Theorem 9 *For any policy π, $\underline{IVI}_\pi$ and $\overline{IVI}_\pi$ are contraction mappings.*

Proof: We first show that $\overline{IVI}_\pi$ is a contraction mapping on $\mathcal{V}$, the space of value functions. Strictly speaking, $\overline{IVI}_\pi$ is a mapping from an interval value function $\hat{V}$ to a value function V. However, the specific values $V(p)$ only depend on the upper bounds $\underline{V}$ of $\hat{V}$. Therefore, the mapping $\overline{IVI}_\pi$ is isomorphic to a function which maps value functions to value functions and

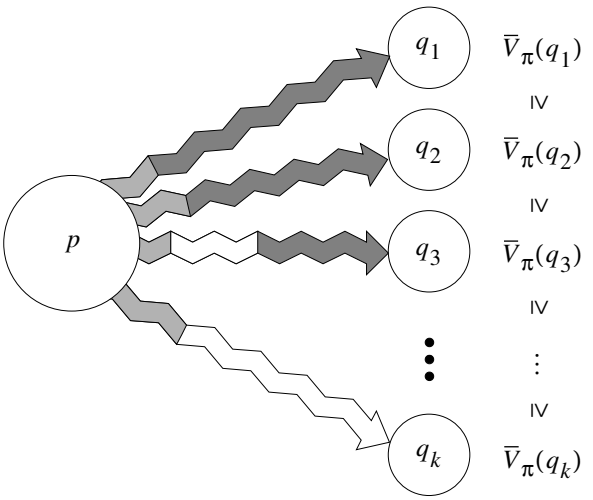

Figure 2: An illustration of the basic dynamic programming step in computing an interval value function for a fixed policy and bounded parameter MDP. The lighter shaded portions of each arc represent the required lower bound transition probability and the darker shaded portions represent the fraction of the remaining transition probability to the upper bound assigned to the arc by F.

with some abuse of terminology, we can consider $\overline{IVI}_\pi$ to be a mapping from $\mathcal{V}$ to $\mathcal{V}$. The same is true for $\underline{IVI}_\pi(\hat{V})$, the value function extracted from the lower bounds of $\hat{V}$.

Let $\hat{u}$ and $\hat{v}$ be interval value functions where $\overline{u}, \overline{v} \in \mathcal{V}$, fix $p \in \mathcal{Q}$ and assume that $\overline{IVI}_\pi(\hat{v})(p) \geq \overline{IVI}_\pi(\hat{u})(p)$. Let F^* be the transition function of the MDP $M \in \mathcal{M}$ that maximizes the following expression involving $\overline{v}$:

$$R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\pi(p))\overline{v}(q).$$

Then,

$$\begin{aligned}
0 &\leq \overline{IVI}_\pi(\hat{v})(p) - \overline{IVI}_\pi(\hat{u})(p) \\
&= \max_{M \in \mathcal{M}} VI_{M,\pi(p)}(\overline{v})(p) - \max_{M \in \mathcal{M}} VI_{M,\pi(p)}(\overline{u})(p) && (3) \\
&\leq R(p) + \gamma \sum_{q \in \mathcal{Q}} F^*_{pq}(\pi(p))\overline{v}(q) - R(p) - \gamma \sum_{q \in \mathcal{Q}} F^*_{pq}(\pi(p))\overline{u}(q) && (4) \\
&= \gamma \sum_{q \in \mathcal{Q}} F^*_{pq}(\pi(p))[\overline{v}(q) - \overline{u}(q)] && (5) \\
&\leq \gamma \sum_{q \in \mathcal{Q}} F^*_{pq}(\pi(p))\|\overline{v} - \overline{u}\| && (6) \\
&= \gamma \|\overline{v} - \overline{u}\|. && (7)
\end{aligned}$$

Line (3) expands the definition of $\overline{IVI}_\pi$. Line (4) follows from the fact

that by definition, $\overline{IVI}_\pi$ chooses the MDP (and the corresponding transition function) that maximizes the value of $VI_{M,\pi(p)}(p)$; any other choice (*cf.* F^*) would result in a lesser value. In Line (5), we simplify the expression by subtracting the immediate reward term and factoring out the coefficients F^*. In Line (fourth), we introduce an inequality by replacing all the terms $[\overline{v}(q) - \overline{u}(q)]$ with the maximum difference over all states, which by definition is the sup norm. The final step (7) follows from the fact that F^* is a probability distribution that sums to 1 and $\|\overline{v} - \overline{u}\|$ does not depend on q.

Repeating this argument in the case that $\overline{IVI}_\pi(\hat{v})(p) \leq \overline{IVI}_\pi(\hat{u})(p)$ implies that

$$|\overline{IVI}_\pi(\hat{v})(p) - \overline{IVI}_\pi(\hat{u})(p)| \leq \gamma\|\overline{v} - \overline{u}\|$$

for all $p \in \mathcal{Q}$. Taking the maximum over p in the above expression gives the result.

The proof that $\underline{IVI}_\pi$ is a contraction mapping is identical, except that $\overline{u}$ and $\overline{v}$ are replaced by $\underline{u}$ and $\underline{v}$, respectively. Also, F^* minimizes the following expression involving $\underline{u}$:

$$R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\pi(p))\underline{u}(q).$$

□

Theorem 10 *For any policy π, $\underline{V}_\pi$ is a fixed-point of $\underline{IVI}_\pi$ and $\overline{V}_\pi$ of $\overline{IVI}_\pi$.*

Proof: We prove the theorem for $\underline{IVI}_\pi$; the proof for $\overline{IVI}_\pi$ is similar. We show

(a) $\underline{IVI}_\pi(\hat{V}_\pi) \leq_{\text{dom}} \underline{V}_\pi$, and

(b) $\underline{IVI}_\pi(\hat{V}_\pi) \geq_{\text{dom}} \underline{V}_\pi$

from which we conclude that $\underline{IVI}_\pi(\hat{V}_\pi) = \underline{V}_\pi$.

We first prove (a). Let M^* be the π-minimizing MDP of

$$\underline{V}_\pi = \min_{M \in \mathcal{M}} V_{M,\pi}$$

which by Theorem 1 exists. From Theorem 5, we know that $V_{M^*,\pi}$ is a fixed point of $VI_{M^*,\pi}(V_{M^*,\pi})$. Thus, for a state $p \in \mathcal{Q}$,

$$\underline{V}_\pi(p) = V_{M^*,\pi}(p) = VI_{M^*,\pi(p)}(V_{M^*,\pi})(p).$$

Using this fact and expanding the definition of $\underline{IVI}_\pi$, we have

$$\begin{aligned} \underline{IVI}_\pi(\hat{V}_\pi)(p) &= \min_{M \in \mathcal{M}} VI_{M,\pi(p)}(\underline{V}_\pi)(p) \\ &\leq VI_{M^*,\pi(p)}(\underline{V}_\pi)(p) = (\underline{V}_\pi)(p). \end{aligned}$$

To prove (b), suppose for sake of contradiction that $\underline{IVI}_\pi(\hat{V}_\pi) <_{\text{dom}} \underline{V}_\pi$. Then for some state p, $\underline{IVI}_\pi(\hat{V}_\pi)(p) < \underline{V}_\pi(p)$. Let M^* be an MDP that achieves the minimum in

$$\min_{M \in \mathcal{M}} VI_{M,\pi(p)}(\underline{V}_\pi)(p).$$

Then, substituting M^* into the definition of $\underline{IVI}_\pi$,

$$\underline{IVI}_\pi(\hat{V}_\pi)(p) = VI_{M^*,\pi(p)}(\underline{V}_\pi)(p) < \underline{V}_\pi(p)$$

However, from Theorem 8, if for a particular MDP M, the result of $VI_{M,\pi(p)}$ applied to a value function v is less than v, the actual value of the MDP is less than v. Therefore, for the MDP M^*,

$$V_{M^*,\pi}(p) < \underline{V}_\pi(p).$$

However, we obtain a contradiction since $\underline{V}_\pi(p)$ is equivalent to the value function of the MDP with the minimal value function over all MDPs in $\mathcal{M}$. □

These theorems, together with Theorem 3 (the Banach fixed-point theorem) imply that iterating $\hat{IVI}_\pi$ on any initial interval value function converges to $\hat{V}_\pi$, regardless of the starting point.

5.2 Interval Value Iteration

As in the case of VI_π and VI, it is straightforward to modify $\hat{IVI}_\pi$ so that it computes optimal policy value intervals by adding a maximization step over the different action choices in each state. However, unlike standard value iteration, the quantities being compared in the maximization step are closed real intervals, so the resulting algorithm varies according to how we choose to compare real intervals. We define two variations of interval value iteration—other variations are possible.

$$\hat{IVI}_{opt}(\hat{V})(p) = \max_{\alpha \in \mathcal{A},\, \leq_{\text{opt}}} \left[\min_{M \in \mathcal{M}} VI_{M,\alpha}(\underline{V})(p),\ \max_{M \in \mathcal{M}} VI_{M,\alpha}(\overline{V})(p) \right]$$

$$\hat{IVI}_{pes}(\hat{V})(p) = \max_{\alpha \in \mathcal{A},\, \leq_{\text{pes}}} \left[\min_{M \in \mathcal{M}} VI_{M,\alpha}(\underline{V})(p),\ \max_{M \in \mathcal{M}} VI_{M,\alpha}(\overline{V})(p) \right]$$

The added maximization step introduces no new difficulties in implementing the algorithm. We discuss convergence for $\hat{IVI}_{opt}$—the convergence results for $\hat{IVI}_{pes}$ are similar. $\overline{IVI}_{opt}$ can be easily shown to be a contraction mapping, and with somewhat more difficulty it can be shown that $\hat{V}_{\text{opt}}$ is a

fixed point of $\hat{IVI}_{opt}$. It then follows that $\overline{IVI}_{opt}$ converges to $\overline{V}_{\rm opt}$. The analogous results for $\underline{IVI}_{opt}$ are somewhat more problematic. Because the action selection is done according to $\leq_{\rm opt}$, which focuses primarily on the interval upper bounds, the mapping $\underline{IVI}_{opt}$ is not a contraction mapping in general. However, once the upper bounds $\overline{IVI}_{opt}$ have converged (*i.e.*, the upper bounds $\overline{V}$ of the input function are a fixed point of $\overline{IVI}$), $\underline{IVI}_{opt}$ acts as a contraction mapping (this statement is formalized below), and the desired convergence will then occur.

Theorem 11 *1. $\overline{IVI}_{opt}$ and $\underline{IVI}_{pes}$ are contraction mappings.*

2. *On value functions $\hat{V}$ for which $\overline{V}$ is a fixed point of $\overline{IVI}_{opt}$, $\underline{IVI}_{opt}$ is a contraction mapping.*

3. *On value functions $\hat{V}$ for which $\underline{V}$ is a fixed point of $\underline{IVI}_{pes}$, $\overline{IVI}_{pes}$ is a contraction mapping.*

Proof: (1) The proof for $\overline{IVI}_{opt}$ follows exactly as in Theorem 5.1 except that we replace the action $\pi(p)$ with the action α^* that maximizes

$$\max_{M \in \mathcal{M}} R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\alpha) v(q).$$

and let F^* be the transition function of the MDP $M \in \mathcal{M}$ that maximizes

$$R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\alpha^*) v(q).$$

Similarly, for $\underline{IVI}_{pes}$, α^* is the action that maximizes

$$\min_{M \in \mathcal{M}} R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\alpha) u(q).$$

and F^* is the transition function of the MDP $M \in \mathcal{M}$ that minimizes

$$R(p) + \gamma \sum_{q \in \mathcal{Q}} F_{pq}(\alpha^*) u(q).$$

(2) TO BE CONTINUED (3) TO BE CONTINUED □

Theorem 12 *$\hat{V_{\rm opt}}$ is a fixed-point of $\hat{IVI}_{opt}$, and $\hat{V_{\rm pes}}$ of $\hat{IVI}_{pes}$.*

Theorem 13 *Iteration of $\overline{IVI}_{opt}$ converges to $\overline{V}_{\rm opt}$, and iteration of $\underline{IVI}_{pes}$ converges to $\underline{V}_{\rm pes}$.*

Proof: By Theorem 11(1), $\overline{IVI}_{opt}$ and $\underline{IVI}_{pes}$ are contraction mappings so that the hypotheses of Theorem 3 are satisfied. Therefore there exists an interval value function $\hat{V}_1$ whose upper bound $\overline{V}_1 \in \mathcal{V}$ is the unique solution to $v = \overline{IVI}_{opt}(v)$ and there exists an interval value function $\hat{V}_2$ whose lower bound $\underline{V}_2 \in \mathcal{V}$ is the unique solution to $u = \underline{IVI}_{pes}(u)$. Theorem 12 establishes that $\overline{V}_1 = \overline{V}_{\text{opt}}$ and $\underline{V}_2 = \underline{V}_{\text{pes}}$. □

We have not yet shown that the other bounds will converge as desired in general. However, this weakness is mitigated by several factors. First, we are far more interested in the values of $\overline{IVI}_{opt}$ and $\underline{IVI}_{pes}$ than in the other bounds—as these represent more useful numbers. $\overline{IVI}_{opt}$ represents the best possible value we could obtain from each state, and $\underline{IVI}_{pes}$ represents the best value we can *guarantee* obtaining from each state (in contrast, for example, $\overline{IVI}_{pes}$ is the best value we might obtain if we follow the pessimistic optimal policy, which was not designed for maximizing this value in any case). Second, we expect to be able to show that these algorithms converge in time polynomial in their input sizes, for fixed discount rate—in that case, we will be able to count on $\overline{IVI}_{opt}$ converging quickly so that $\underline{IVI}_{opt}$ will then converge. Third, we can use $\overline{IVI}_{opt}$ or $\underline{IVI}_{pes}$ to select an optimal policy, and then for that policy, we can use interval policy evaluation to get both lower and upper bounds for its value.

6 Policy Selection, Sensitivity Analysis, and Aggregation

In this section, we consider some basic issues concerning the use and interpretation of bounded parameter MDPs. We begin by reemphasizing some ideas introduced earlier regarding the selection of policies.

To begin with, it is important that we are clear on the status of the bounds in a bounded parameter MDP. A bounded parameter MDP specifies upper and lower bounds on individual parameters; the assumption is that we have no additional information regarding individual exact MDPs whose parameters fall with those bounds. In particular, we have no prior over the exact MDPs in the family of MDPs defined by a bounded parameter MDP.

Policy selection Despite the lack of information regarding any particular MDP, we may have to choose a policy. In such a situation, it is natural to consider that the actual MDP, *i.e.,* the one in which we will ultimately have to carry out some policy, is decided by some outside process. That process might choose so as to help or hinder us, or it might be entirely indifferent. To minimize the risk of performing poorly, it is reasonable to think in adversarial terms; we select the policy which will perform as well as possible assuming

that the adversary chooses so that we perform as poorly as possible.

These choices correspond to optimistic and pessimistic optimal policies. We have discussed in the last section how to compute interval value functions for such policies—such value functions can then be used in a straightforward manner to extract policies which achieve those values.

There are other possible choices, corresponding in general to other means of totally ordering real closed intervals. We might for instance consider a policy whose average performance over all MDPs in the family is as good as or better than the average performance of any other policy. This notion of average is potentially problematic, however, as it essentially assumes a uniform prior over exact MDPs and, as stated earlier, the bounds do not imply any particular prior.

Sensitivity analysis There are other ways in which bounded parameter MDPs might be useful in planning under uncertainty. For example, we might assume that we begin with a particular exact MDP, say, the MDP with parameters whose values reflect the best guess according to a given domain expert. If we were to compute the optimal policy for this exact MDP, we might wonder about the degree to which this policy is sensitive to the numbers supplied by the expert.

To explore this possible sensitivity to the parameters, we might assess the policy by perturbing the parameters and evaluating the policy with respect to the perturbed MDP. Alternatively, we could use bounded parameter MDPs to perform this sort of sensitivity analysis on a whole family of MDPs by converting the point estimates for the parameters to confidence intervals and then computing bounds on the value function for the fixed policy via interval policy evaluation.

Aggregation Another use of BMDPs involves a different interpretation altogether. Instead of viewing the states of the bounded parameter MDP as individual primitive states, we view each state of the BMDP as representing a set or *aggregate* of states of some other, larger MDP.

In this interpretation, states are aggregated together because they behave approximately the same with respect to possible state transitions. A little more precisely, suppose that the set of states of the BMDP $\mathcal{M}$ corresponds to the set of *blocks* $\{B_1, \ldots, B_n\}$ such that the $\{B_i\}$ constitutes the partition of another MDP with a much larger state space.

Now we interpret the bounds as follows; for any two blocks B_i and B_j, let $\hat{F}_{B_iB_j}(\alpha)$ represent the interval value for the transition from B_i to B_j on action α defined as follows:

$$\hat{F}_{B_iB_j}(\alpha) = \left[\min_{p \in B_i} \sum_{q \in B_j} F_{pq}(\alpha), \ \max_{p \in B_i} \sum_{q \in B_j} F_{pq}(\alpha) \right]$$

Intuitively, this means that all states in a block behave approximately the same (assuming the lower and upper bounds are close to each other) in terms of transitions to other blocks even though they may differ widely with regard to transitions to individual states.

Dean *et al.* [1997] discuss methods for using an implicit representation of a exact MDP with a large number of states to construct a BMDP with a much smaller number of states based on an aggregation like that just discussed. We then show how to use this BMDP to compute policies with desirable properties with respect to the large implicitly described MDP. Note that the original implicit MDP is not even a member of the family of MDPs for the reduced BMDP (it has a different state space, for instance). Nevertheless, it is a theorem that the policies and value bounds of the BMDP can be soundly applied in the original MDP (using the aggregation mapping to connect original states to the block-states of the BMDP).

7 Related Work and Conclusions

Our definition for bounded parameter MDPs is related to a number of other ideas appearing in the literature on Markov decision processes; in the following, we mention just a few such ideas. Bertsekas and Castañon [1989] use the notion of aggregated Markov chains and consider grouping together states with approximately the same residuals. Methods for bounding value functions are frequently used in approximate algorithms for solving MDPs; Lovejoy [1991] describes their use in solving partially observable MDPs. Puterman [Puterman, 1994] provides an excellent introduction to Markov decision processes and estimation techniques that involve bounding value functions.

Boutilier and Dearden [1994] and Boutilier *et al.* [1995b] describe methods for solving implicitly described MDPs and Dean and Givan [1997] reinterpret this work in terms of computing explicitly described MDPs with aggregate states. Littman [1996] provides a generalization of MDPs that may encompass some aspects of the theory of bounded parameter MDPs.

Bounded parameter MDPs allow us to represent uncertainty or variation in the parameters of a target Markov decision process. Interval value functions capture the variation in expected value for policies as a consequence of variation in parameter estimates. In this paper, we have precisely defined the notions of bounded parameter MDP and interval value function, provided algorithms for computing interval value functions, and described methods and algorithms for policy selection and evaluation, sensitivity analysis, and aggregation.

References

[Bellman, 1957] Bellman, Richard 1957. *Dynamic Programming.* Princeton University Press.

[Bertsekas and Castañon, 1989] Bertsekas, D. P. and Castañon, D. A. 1989. Adaptive aggregation for infinite horizon dynamic programming. *IEEE Transactions on Automatic Control* 34(6):589–598.

[Boutilier and Dearden, 1994] Boutilier, Craig and Dearden, Richard 1994. Using abstractions for decision theoretic planning with time constraints. In *Proceedings AAAI-94.* AAAI. 1016–1022.

[Boutilier *et al.*, 1995a] Boutilier, Craig; Dean, Thomas; and Hanks, Steve 1995a. Planning under uncertainty: Structural assumptions and computational leverage. In *Proceedings of the Third European Workshop on Planning.*

[Boutilier *et al.*, 1995b] Boutilier, Craig; Dearden, Richard; and Goldszmidt, Moises 1995b. Exploiting structure in policy construction. In *Proceedings IJCAI 14.* IJCAII. 1104–1111.

[Dean and Givan, 1997] Dean, Thomas and Givan, Robert 1997. Model minimization in Markov decision processes. In *Proceedings AAAI-97.* AAAI.

[Dean *et al.*, 1997] Dean, Thomas; Givan, Robert; and Leach, Sonia 1997. Model reduction techniques for computing approximately optimal solutions for Markov decision processes. In *Thirteenth Conference on Uncertainty in Artificial Intelligence.*

[Littman *et al.*, 1995] Littman, Michael; Dean, Thomas; and Kaelbling, Leslie 1995. On the complexity of solving Markov decision problems. In *Eleventh Conference on Uncertainty in Artificial Intelligence.* 394–402.

[Littman, 1996] Littman, Michael Lederman 1996. *Algorithms for Sequential Decision Making.* Ph.D. Dissertation, Department of Computer Science, Brown University.

[Littman, 1997] Littman, Michael L. 1997. Probabilistic propositional planning: Representations and complexity. In *Proceedings AAAI-97.* AAAI.

[Lovejoy, 1991] Lovejoy, William S. 1991. A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research* 28:47–66.

[Puterman, 1994] Puterman, Martin L. 1994. *Markov Decision Processes.* John Wiley & Sons, New York.