ELSEVIER

Artificial Intelligence 122 (2000) 71–109

Artificial Intelligence

www.elsevier.com/locate/artint

Bounded-parameter Markov decision processes

Robert Givan [a,*], Sonia Leach [b], Thomas Dean [b]

[a] *Department of Electrical and Computer Engineering, Purdue University, 1285 EE Building, West Lafayette, IN 47907, USA*

[b] *Department of Computer Science, Brown University, 115 Waterman Street, Providence, RI 02912, USA*

Received 28 May 1999; received in revised form 22 May 2000

Abstract

In this paper, we introduce the notion of a *bounded-parameter Markov decision process* (BMDP) as a generalization of the familiar *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). BMDPs form an efficiently solvable special case of the already known class of MDPs with *imprecise parameters* (MDPIPs). Bounded-parameter MDPs can be used to represent variation or uncertainty concerning the parameters of sequential decision problems in cases where no prior probabilities on the parameter values are available. 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* that computes an interval value function for a given BMDP 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. We define *optimistic* and *pessimistic* criteria for optimality, and provide a variant of value iteration (Bellman, 1957) that we call *interval value iteration* that computes policies for a BMDP that are optimal with respect to these criteria. We show that each algorithm we present converges to the desired values in a polynomial number of iterations given a fixed discount factor.

* Corresponding author.

E-mail addresses: givan@ecn.purdue.edu (R. Givan), sml@cs.brown.edu (S. Leach), tld@cs.brown.edu (T. Dean).

0004-3702/00/$ – see front matter
PII: S0004-3702(00)00047-3

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

1. Introduction

The theory of Markov decision processes (MDPs) [1,2,10,11,14] provides the semantic foundations for a wide range of problems involving planning under uncertainty [5,7]. Most work in the planning subarea of artificial intelligence addresses problems that can be formalized using MDP models—however, it is often the case that such models are exponentially larger than the original "intensional" problem representation used in AI work. This paper generalizes the theory of MDPs in a manner that is useful for more compactly representing AI problems as MDPs via state-space aggregation, as we discuss below.

In this paper, we introduce a generalization of Markov decision processes called *bounded-parameter Markov decision processes* (BMDPs) that allows us to model uncertainty about 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. While any policy on the original (primitive) states induces a stationary distribution over those states that can be used to give prior probabilities to the different transition probabilities in the intervals, we may be unable to compute these prior probabilities—the original reason for aggregating the states is typically to avoid such expensive computation over the original large state space.

Aggregation of states in very large state spaces was our original motivation for developing BMDPs. Substantial effort has been devoted in recent years within the AI community [6,8,9] to the problem of representing and reasoning with MDP problems where the state space is not explicitly listed but rather implicitly specified with a *factored representation*. In such problems, an explicit listing of the possible system states is exponentially longer than the more natural implicit problem description, and such an explicit list is often intractable to work with. Most planning problems of interest to AI researchers fit this description in that they are only representable in reasonable space using implicit representations. Recent work in applying MDPs to such problems (e.g., [6,8,9]) has considered state-space aggregation techniques as a means of dealing with this problem: rather than work with the possible system states explicitly, aggregation techniques work with blocks of similar or identically-behaving states. When aggregating states that have similar but not identical behavior, the question immediately arises of what transition probability holds between the aggregates: this probability will depend on which underlying state is in control, but this choice of underlying state is not modelled in the aggregate

model. This work can be viewed as providing a means of addressing this problem by allowing intervals rather than point values for the aggregate transition probabilities: the interval can be chosen to include the true value for each of the underlying states present in the aggregates involved. It should be noted that under these circumstances, deriving a prior probability distribution over the true parameter values is often as expensive as simply avoiding the aggregation altogether and would defeat the purpose entirely. Moreover, assuming any particular probability distribution could produce arbitrarily inaccurate results. As a result, this work considers parameters falling into intervals with no prior probability distribution specified over the possible parameter values in the intervals, and seeks to put bounds on how badly or how well particular plans will perform in such a context, as well as to provide means to find optimal plans under optimistic or pessimistic assumptions about the true distribution over parameter values. In Section 6, we discuss the application of our BMDP approach to state-space aggregation problems more formally. Also, in a related paper, we have shown how BMDPs can be used as part of an state-space aggregation strategy for efficiently approximating the solution of MDPs with very large state spaces and dynamics compactly encoded in a factored (or implicit) representation [10].

We also discuss later in this paper the potential use of BMDP methods to evaluate the sensitivity of the optimal policy in an exact MDP to small variations in the parameter values defining the MDP—using BMDP policy selection algorithms on a BMDP whose parameter intervals represent small variations (perhaps confidence intervals) around the exact MDP parameter values, the best and worst variation in policy value achieved can be measured.

In this paper we introduce and discuss BMDPs, the BMDP analog of value functions, called *interval value functions*, and policy selection and evaluation methods for BMDPs. 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 optimal value functions 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 in polynomial time, for a fixed discount factor. Finally, we consider two different notions of optimal policy for a 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.

Our interval policy evaluation and interval value iteration algorithms rely on iterative convergence to the desired values, and are generalizations of the standard MDP algorithms *successive approximation* and *value iteration*, respectively. We believe it is also possible to design an interval-valued variant of the standard MDP algorithm *policy iteration*, but we have not done so at this writing—however, it should be clear that our successive approximation algorithm for evaluating policies in the BMDP setting provides an essential basic building block for constructing a policy iteration method; all that need be added is a means for selecting a new action at each state based on the interval value function of the preceding policy (and a possibly difficult corresponding analysis of the properties of the algorithm). We note that there is no consensus in the decision-theoretic planning and learning and operations-research communities as to whether value iteration, policy

iteration, or even standard linear programming is generally the best approach to solving MDP problems: each technique appears to have its strengths and weaknesses.

BMDPs are an efficiently solvable specialization of the already known class of *Markov Decision Processes with Imprecisely Known Transition Probabilities* (MDPIPs) [15,17, 18]. In the related work section we discuss in more detail how BMDPs relate to MDPIPs.

Here is a high-level overview of how conceptual, theoretical, algorithmic, and experimental treatments are woven together in the remainder of the paper. We begin by introducing the concept of a Bounded-Parameter MDP (BMDP), and introducing and justifying BMDP analogues for optimal policies and value functions. In terms of the theoretical development, we define the basic mathematical objects, introduce notational conventions, and provide some background in MDPs. We define the objects and operations that will be useful in the subsequent theoretical and algorithmic development, e.g., composition operators on MDPs and on policies. Finally, we define and motivate the relevant notions of optimality, and then prove the existence of optimal policies with respect to the different notions of optimality.

In addition to this theoretical and conceptual development, in terms of algorithm development we describe and provide pseudocode for algorithms for computing optimal policies and value functions with respect to the different notions of optimality, e.g., interval policy evaluation and interval value iteration. We provide an analysis of the complexity of these algorithms and prove that they compute optimal policies as defined earlier. We then describe a proof-of-concept implementation and summarize preliminary experimental results. We also provide a brief overview of some applications including sensitivity analysis, coping with parameters known to be imprecise, and support for state aggregation methods. Finally, we survey some additional related work not covered in the primary text and summarize our contributions.

Before introducing BMDPs and their algorithms in Section 4 and Section 5, we first present in the next two sections a brief review of exact MDPs, policy evaluation, and value iteration in order to establish notational conventions we use throughout the paper. Our presentation follows that of [14], where a more complete account may be found.

2. Exact Markov decision processes

An (exact) Markov decision process M is a four-tuple $M = \langle Q, A, F, R\rangle$ where Q is a set of states, 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 $\alpha \in A$ and $p, q \in Q$

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

where X_t and U_t are random variables denoting, respectively, the state and action at time t. When needed we write F^M to denote the transition function of the MDP M.

A *policy* is a mapping from states to actions, $\pi : Q \to A$. The set of all policies is denoted Π. An MDP M together with a fixed policy $\pi \in \Pi$ determines a Markov chain

[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.

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

$$V_{M,\pi}(p) = R(p) + \gamma \sum_{q \in Q} F_{pq}\big(\pi(p)\big) V_{M,\pi}(q), \tag{2}$$

where $0 \leqslant \gamma < 1$ 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.

$$V^*(p) = \max_{\alpha \in A} \left(R(p) + \gamma \sum_{q \in Q} F_{pq}(\alpha) V^*(q) \right). \tag{3}$$

The value function V^* is greater than or equal to any value function V_π in the partial order $\geqslant_{\text{dom}}$ defined as follows: $V_1 \geqslant_{\text{dom}} V_2$ if and only if for all states q, $V_1(q) \geqslant V_2(q)$ (in this case we say that V_1 *dominates* V_2). We write $V_1 >_{\text{dom}} V_2$ to mean $V_1 \geqslant_{\text{dom}} V_2$ and for at least one state q, $V_1(q) > V_2(q)$.

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 in the definition of V^* with argmax.

3. 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. We follow the example of [14]. In Section 5, 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 that 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 1) to show that iterating these operators converges to unique fixed points, V_π and V^* respectively (Theorems 3 and 4).

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

$$\|v\| = \max_{q \in Q} |v(q)|. \tag{4}$$

We use the term *convergence* to mean convergence in the norm sense. The space $\overline{V}$ 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, $0 \leqslant \lambda < 1$,

[2] In this paper, we focus on expected discounted cumulative reward as a performance criterion, but other criteria, e.g., total or average reward [14], are also applicable to bounded-parameter MDPs.

such that $\|Tv - Tu\| \leqslant \lambda\|v - u\|$ for all u and v in U. Define $VI : \overline{V} \to \overline{V}$ and for each $\pi \in \Pi$, $VI_\pi : \overline{V} \to \overline{V}$ on each $p \in Q$ by

$$VI(v)(p) = \max_{\alpha \in A}\left(R(p) + \gamma \sum_{q \in Q} F_{pq}(\alpha)v(q) \right), \quad \text{and} \tag{5}$$

$$VI_\pi(v)(p) = R(p) + \gamma \sum_{q \in Q} F_{pq}\big(\pi(p)\big)v(q). \tag{6}$$

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

$$\begin{aligned} VI_{M,\pi}(v)(p) &= R(p) + \gamma \sum_{q \in Q} F^M_{pq}\big(\pi(p)\big)v(q), \\ VI_{M,\alpha}(v)(p) &= R(p) + \gamma \sum_{q \in Q} F^M_{pq}(\alpha)v(q). \end{aligned} \tag{7}$$

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

$$V^*(p) = VI(V^*)(p) \quad \text{and} \quad V_\pi(p) = VI_\pi(V_\pi)(p) \tag{8}$$

for all states $p \in Q$. 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. Proofs for these theorems can be found in the work of Puterman [14].

Theorem 1. *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 2. *VI and VI_π are contraction mappings.*

Theorems 1 and 2 together prove the following fundamental results in the theory of MDPs.

Theorem 3. *There exists a unique $v^* \in \overline{V}$ satisfying $v^* = VI(v^*)$; furthermore, $v^* = V^*$. Similarly V_π is the unique fixed point of VI_π.*

Theorem 4. *For arbitrary $v^0 \in \overline{V}$, 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 4 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 $\overline{V}$, and repeatedly apply the operator VI to obtain the sequence $\{v^n\}$. This algorithm is referred to as *value iteration*. Theorem 4 guarantees the convergence of value iteration to

the optimal value function. Similarly, we can specify an algorithm called *policy evaluation* that finds V_π by repeatedly applying VI_π starting with an initial $v^0 \in \overline{V}$.

The following theorem from [12] states a convergence rate of value iteration and policy evaluation that can be derived using bounds on the precision needed to represent solutions to a linear program of limited precision (each algorithm can be viewed somewhat nontrivially as solving a linear program).

Theorem 5. *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 is used extensively in the proofs of the succeeding sections results directly from the monotonicity of the VI_π operator with respect to the $\leqslant_{\text{dom}}$ and $\geqslant_{\text{dom}}$ orderings, together with the above theorems.

Theorem 6. *Let $\pi \in \Pi$ be a policy and M an MDP. Suppose there exists $u \in \overline{V}$ for which $u \leqslant_{\text{dom}} (\geqslant_{\text{dom}}) VI_{M,\pi}(u)$, then $u \leqslant_{\text{dom}} (\geqslant_{\text{dom}}) V_{M,\pi}$. Likewise for the orderings $<_{\text{dom}}$ and $>_{\text{dom}}$.*

4. Bounded-parameter Markov decision processes

A *bounded-parameter MDP* (*BMDP*) is a four-tuple $M_\updownarrow = \langle Q, A, F_\updownarrow, R_\updownarrow \rangle$ where Q and A are defined as for MDPs, and $F_\updownarrow$ and $R_\updownarrow$ are analogous to the MDP F and R but yield closed real intervals instead of real values. That is, for any action α and states p, q, $R_\updownarrow(p)$ and $F_{\updownarrow,p,q}(\alpha)$ are both closed real intervals of the form $[l, u]$ for real numbers l and u with $l \leqslant u$, where in the case of $F_\updownarrow$ we require $0 \leqslant l \leqslant u \leqslant 1$.[3] To ensure that $F_\updownarrow$ admits only well-formed transition functions, we require that for any action α and state p, the sum of the lower bounds of $F_{\updownarrow,p,q}(\alpha)$ 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. Fig. 1 depicts the state-transition diagram for a simple BMDP with three states and one action. We use a one-action BMDP to illustrate various concepts in this paper because multi-action systems are awkward to draw, and one action suffices to illustrate the concepts. Note that a one action BMDP or MDP has only one policy available (select the only action at all states), and so represents a trivial control problem.

A BMDP $M_\updownarrow = \langle Q, A, F_\updownarrow, R_\updownarrow \rangle$ defines a set of exact MDPs that, by abuse of notation, we also call $M_\updownarrow$. For any exact MDP $M = \langle Q', A', F', R' \rangle$, we have $M \in M_\updownarrow$ if $Q = Q'$, $A = A'$, and for any action α and states p, q, $R'(p)$ is in the interval $R_\updownarrow(p)$ and $F'_{p,q}(\alpha)$ is in the interval $F_{\updownarrow,p,q}(\alpha)$. We rely on context to distinguish between the tuple view of $M_\updownarrow$ and the set of exact MDPs view of $M_\updownarrow$. In the remaining definitions in this section, the BMDP $M_\updownarrow$ is implicit. Fig. 3 shows an example of an exact MDP belonging to the family

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

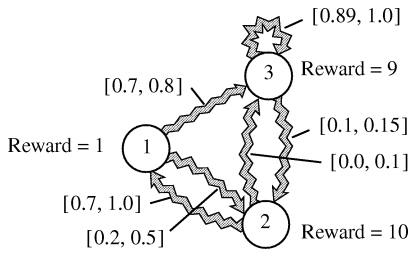

Fig. 1. The state-transition diagram for a simple bounded-parameter Markov decision process with three states and a single action. The arcs indicate possible transitions and are labeled by their lower and upper bounds.

described by the BMDP in Fig. 1. We use the convention that thick wavy lines represent interval valued transition probabilities and thinner straight lines represent exact transition probabilities.

An *interval value function* $V_\updownarrow$ is a mapping from states to closed real intervals. We generally use such functions to indicate that the value of a given state falls within the selected interval. Interval value functions can be specified for both exact MDPs 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 interval valued function (e.g., $F_\updownarrow$, $R_\updownarrow$, $V_\updownarrow$, and those we define later) we define two real valued functions that take the same arguments and return the upper and lower interval bounds, respectively, denoted by the following syntactic variations: $F_\uparrow$, $R_\uparrow$, $V_\uparrow$ for upper bounds, and $F_\downarrow$, $R_\downarrow$, $V_\downarrow$ for lower bounds, respectively. So, for example, at any state q we have $V_\updownarrow(q) = [V_\downarrow(q), V_\uparrow(q)]$.

We note that the number of MDPs $M \in M_\updownarrow$ is in general uncountable. We start our analysis by showing that there is a finite subset $X_{M_\updownarrow} \in M_\updownarrow$ of these MDPs of particular interest. Given any ordering O of all the states in Q, there is a unique MDP $M \in M_\updownarrow$ that minimizes, for every state q and action α, the expected "position in the ordering" of the state reached by taking action α in state q—in other words, an MDP that for every state q and action α sends as much probability mass as possible to states early in the ordering O when taking action α in state q. Formally, we define the following concept:

Definition 1. Let $O = q_1, q_2, \ldots, q_k$ be an ordering of Q. We define the *order-maximizing MDP* M_O with respect to ordering O as follows.

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

$$\sum_{i=1}^{r-1} F_{\uparrow\, p,q_i}(\alpha) + \sum_{i=r}^{k} F_{\downarrow\, p,q_i}(\alpha). \tag{9}$$

The value r is the index into the state ordering $\{q_i\}$ such that below index r we assign the upper bound, and above index r we assign the lower bound, with the rest of the probability

mass from p under α being assigned to q_r. Formally, we select $M_O \in M_{\updownarrow}$ by choosing $F^{M_O}_{p,q}(\alpha)$ for all $q \in Q$ as follows:

$$F^{M_O}_{p,q_j}(\alpha) = \begin{cases} F_{\uparrow pq_i}(\alpha) & \text{if } j < r, \\ F_{\downarrow pq_i}(\alpha) & \text{if } j > r, \end{cases} \quad \text{and}$$

$$F^{M_O}_{p,q_r}(\alpha) = 1 - \sum_{i=1,\, i\neq r}^{i=k} F^{M_O}_{p,q_i}(\alpha).$$

Fig. 2 shows a diagrammatic representation of the order-maximizing MDP at a particular state p for the particular ordering of the state space shown. Fig. 3 shows the order-maximizing MDP for the particular BMDP shown in Fig. 1 using a particular state order $(2 > 3 > 1)$, as a concrete example.

Definition 2. Let $X_{M_{\updownarrow}}$ be the set of order-maximizing MDPs M_O in $M_{\updownarrow}$, one for each ordering O. Note that since there are finitely many orderings of states, $X_{M_{\updownarrow}}$ is finite.

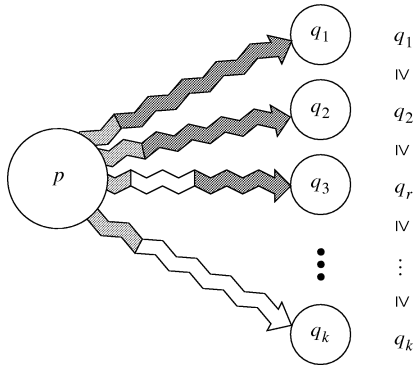

Fig. 2. An illustration of the transition probabilities in the order-maximizing MDP at the state p for the order shown. The lighter shaded portions of each arc represent the required lower bound transition probability and the darker shaded portions represent the fraction of the remaining allowed transition probability assigned to the arc by T.

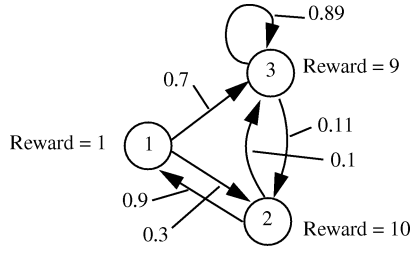

Fig. 3. The order-maximizing MDP for the BMDP shown in Fig. 1 using the state order $2 > 3 > 1$.

We now show that the set $X_{M_\updownarrow}$ in some sense contains every MDP of interest from $M_\updownarrow$. In particular, we show that for any policy π and any MDP M in $M_\updownarrow$, the value of π in M is bracketed by values of π in two MDPs in $X_{M_\updownarrow}$.

Lemma 1. *For any MDP $M \in M_\updownarrow$:*

(a) *For any policy $\pi \in \Pi$, there are MDPs $M_1 \in X_{M_\updownarrow}$ and $M_2 \in X_{M_\updownarrow}$ such that*

$$V_{M_1,\pi} \leqslant_{\mathrm{dom}} V_{M,\pi} \leqslant_{\mathrm{dom}} V_{M_2,\pi}. \tag{10}$$

(b) *Also, for any value function $v \in \overline{V}$, there are MDPs $M_3 \in X_{M_\updownarrow}$ and $M_4 \in X_{M_\updownarrow}$ such that*

$$VI_{M_3,\pi}(v) \leqslant_{\mathrm{dom}} VI_{M,\pi}(v) \leqslant_{\mathrm{dom}} VI_{M_4,\pi}(v). \tag{11}$$

Proof. See Appendix A. □

Interval value functions for policies

We now define the interval analogue to the traditional MDP policy-specific value function V_π, and state and prove some of the properties of this interval value function. The development here requires some care, as one desired property of the definition is not immediate. We first observe that we would like an interval-valued function over the state space that satisfies a Bellman equation like that for traditional MDPs (as given by Eq. (2)). Unfortunately, stating a Bellman equation requires us to have specific transition probability distributions F rather than a range of such distributions. Instead of defining policy value via a Bellman equation, we define the interval value function directly, at each state, as giving the range of values that could be attained at that state for the various choices of F allowed by the BMDP. We then show that the desired minimum and maximum values can be achieved independent of the state, so that the upper and lower bound value functions are just the values of the policy in particular "minimizing" and "maximizing" MDPs in the BMDP. This fact enables the use of the Bellman equations for the minimizing and maximizing MDPs to give an iterative algorithm that converges to the desired values, as presented in Section 5.

Definition 3. For any policy π and state q, we define the *interval value* $V_{\updownarrow\pi}(q)$ of π at q to be the interval

$$V_{\updownarrow\pi}(q) = \Big[\min_{M\in M_\updownarrow} V_{M,\pi}(q), \max_{M\in M_\updownarrow} V_{M,\pi}(q)\Big]. \tag{12}$$

We note that the existence of these minimum and maximum values follows from Lemma 1 and the finiteness of the set $X_{M_\updownarrow}$—because Lemma 1 implies that $V_{\updownarrow\pi}(q)$ is the same as the following where the minimization and maximization are done over finite sets:

$$V_{\updownarrow\pi}(q) = \Big[\min_{M\in X_{M_\updownarrow}} V_{M,\pi}(q), \max_{M\in X_{M_\updownarrow}} V_{M,\pi}(q)\Big]. \tag{13}$$

In preparation for the discussion in Section 5, we show in Theorem 7 that for any policy there is at least one specific *policy-maximizing* MDP in $M_\updownarrow$ that achieves the upper bound in Definition 3 at all states q simultaneously (and likewise a different specific *policy-*

minimizing MDP that achieves the lower bound at all states q simultaneously). We formally define these terms below.

Definition 4. For any policy π, an MDP $M \in M_{\updownarrow}$ is π*-maximizing* if $V_{M,\pi}$ dominates $V_{M',\pi}$ for any $M' \in M_{\updownarrow}$, i.e., for any $M \in M'_{\updownarrow}$, $V_{M,\pi} \geqslant_{\text{dom}} V_{M',\pi}$. Likewise, $M \in M_{\updownarrow}$ is π*-minimizing* if it is dominated by all such $V_{M',\pi}$, i.e., for any $M' \in M_{\updownarrow}$, $V_{M,\pi} \leqslant_{\text{dom}} V_{M',\pi}$.

Fig. 4 shows the interval value function for the only policy available in the (trivial) one-action BMDP shown in Fig. 1, along with the π-maximizing and π-minimizing MDPs for that policy.

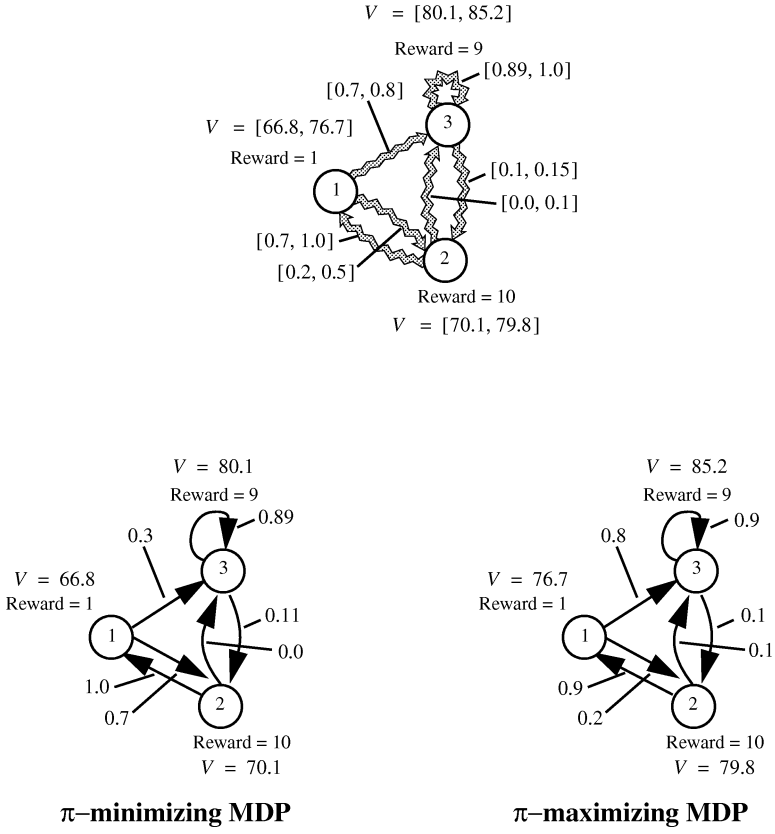

Fig. 4. The interval value function (shown as $V_{\updownarrow}$ on the top subfigure), policy-minimizing MDP with state values (lower left), and policy-maximizing MDP with state values (lower right) for the one-action BMDP shown in Fig. 1 under the only policy. We assume a discount factor of 0.9. Note that the lower-bound values in the interval value function are the state values under the policy-minimizing MDP, and the upper-bound values are the state values under the policy-maximizing MDP. Also, note that the policy-maximizing MDP is the order-maximizing MDP for the state order $3 > 2 > 1$ and the policy-minimizing MDP is the order-maximizing MDP for the order $1 > 2 > 3$—policy-minimizing and -maximizing MDPs are always order-maximizing for some order (but the orders need not be reverse to one another as they are in this example).

We note that Lemma 1 implies that for any single state q and any policy π we can select an MDP $M \in M_{\updownarrow}$ to maximize (or minimize) $V_{M,\pi}(q)$ by selecting the MDP in $X_{M_{\updownarrow}}$ that gives the largest value for π at q. However, we have not shown that a single MDP can be chosen to simultaneously maximize (or minimize) $V_{M,\pi}(q)$ at all states $q \in Q$ (i.e., that there exist π-maximizing and π-minimizing MDPs). In order to show this fact, we show how to compose two MDPs (with respect to a fixed policy π) to construct a third MDP such that the value of π in the third MDP is not less than the value of π in either of the initial two MDPs, at every state. We can then construct a π-maximizing MDP by composing together all the MDPs that maximize the value of π at the different individual states (likewise for π-minimizing MDPs using a similar composition operator). We start by defining the just mentioned policy-relative composition operators on MDPs:

Definition 5. Let $\oplus^{\pi}_{\max}$ and $\oplus^{\pi}_{\min}$ denote composition operators on MDPs with respect to a policy $\pi \in \Pi$, defined as follows:

If $M_1, M_2 \in M_{\updownarrow}$, then $M_3 = M_1 \oplus^{\pi}_{\max} M_2$ if for all states $p, q \in Q$,

$$F^{M_3}_{pq}(\alpha) = \begin{cases} F^{M_1}_{pq}(\alpha) & \text{if } V_{M_1,\pi}(p) \geqslant V_{M_2,\pi}(p) \text{ and } \alpha = \pi(p), \\ F^{M_2}_{pq}(\alpha) & \text{otherwise.} \end{cases}$$

If $M_1, M_2 \in M_{\updownarrow}$, then $M_3 = M_1 \oplus^{\pi}_{\min} M_2$ if for all states $p, q \in Q$,

$$F^{M_3}_{pq}(\alpha) = \begin{cases} F^{M_1}_{pq}(\alpha) & \text{if } V_{M_1,\pi}(p) \leqslant V_{M_2,\pi}(p) \text{ and } \alpha = \pi(p), \\ F^{M_2}_{pq}(\alpha) & \text{otherwise.} \end{cases}$$

We give as an example in Fig. 5 two MDPs from the BMDP of Fig. 1, along with their composition under the $\oplus^{\pi}_{\max}$ operator where π is the single available policy for that

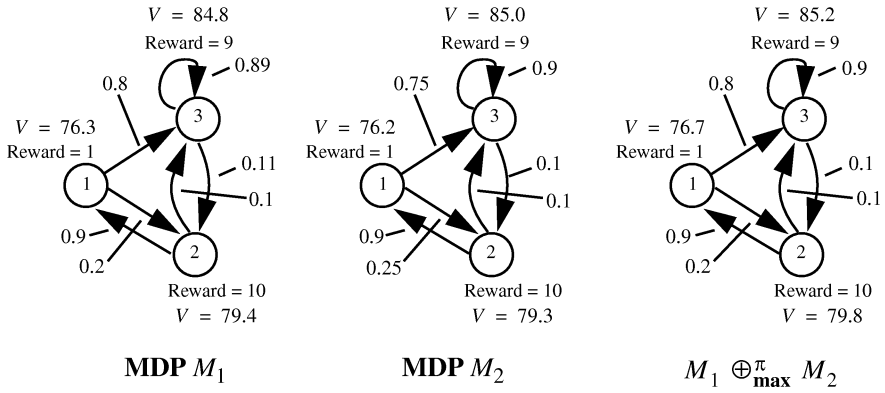

Fig. 5. Two MDPs M_1 and M_2 from the BMDP shown in Fig. 1, and their composition under $\oplus^{\pi}_{\max}$ where π is the only available policy in the one-action BMDP. State-transition probabilities for the composition MDP are selected from the component MDP that achieves the greater value for the source state of the transition. State values are shown for all three MDPs—note that the composition MDP achieves higher value at every state, as claimed in Lemma 2.

one-action BMDP. We now state the property claimed above for this MDP composition operator:

Lemma 2. *Let π be a policy in Π and M_1, M_2 be MDPs in $M_{\updownarrow}$.*

(a) *For $M_3 = M_1 \oplus^{\pi}_{\max} M_2$,*

$$V_{M_3,\pi} \geqslant_{\mathrm{dom}} V_{M_1,\pi} \quad \textit{and} \quad V_{M_3,\pi} \geqslant_{\mathrm{dom}} V_{M_2,\pi}, \quad \textit{and} \tag{14}$$

(b) *for $M_3 = M_1 \oplus^{\pi}_{\min} M_2$,*

$$V_{M_3,\pi} \leqslant_{\mathrm{dom}} V_{M_1,\pi} \quad \textit{and} \quad V_{M_3,\pi} \leqslant_{\mathrm{dom}} V_{M_2,\pi}. \tag{15}$$

Proof. See Appendix A. □

These MDP composition operators can now be used to show the existence of policy-maximizing and policy-minimizing MDPs within $M_{\updownarrow}$.

Theorem 7. *For any policy $\pi \in \Pi$, there exist π-maximizing and π-minimizing MDPs in $X_{M_{\updownarrow}} \subseteq M_{\updownarrow}$.*

Proof. Enumerate $X_{M_{\updownarrow}}$ as a finite sequence of MDPs $M_1, \ldots, M_k$. Consider composing these MDPs together to construct the MDP M as follows:

$$M = \big(\big((M_1 \oplus^{\pi}_{\max} M_2) \oplus^{\pi}_{\max} \cdots\big) \oplus^{\pi}_{\max} M_k\big). \tag{16}$$

Note that M may depend on the ordering of $M_1, \ldots, M_k$, but that any ordering is satisfactory for this proof. It is straightforward to show by induction using Lemma 2 that $V_{M,\pi} \geqslant_{\mathrm{dom}} V_{M_i,\pi}$ for each $1 \leqslant i \leqslant k$, and then Lemma 1 implies that $V_{M,\pi} \geqslant_{\mathrm{dom}} V_{M',\pi}$ for any $M' \in M_{\updownarrow}$. M is thus a π-maximizing MDP. Although M may not be in $X_{M_{\updownarrow}}$, Lemma 1 implies that $V_{M,\pi}$ must be dominated by $V_{M',\pi}$ for some $M' \in X_{M_{\updownarrow}}$, which must also be π-maximizing.

An identical proof implies the existence of π-minimizing MDPs, replacing each occurrence of "max" with "min" and each $\geqslant_{\mathrm{dom}}$ with $\leqslant_{\mathrm{dom}}$. □

Corollary 1. *$V_{\downarrow_\pi} = \min_{M \in M_{\updownarrow}}(V_{M,\pi})$ and $V_{\uparrow_\pi} = \max_{M \in M_{\updownarrow}}(V_{M,\pi})$ where the minimum and maximum are computed relative to $\leqslant_{\mathrm{dom}}$ and are well-defined by Theorem 7.*

We give an algorithm in Section 5 that converges to $V_{\downarrow_\pi}$ by also converging to a π-minimizing MDP in $M_{\updownarrow}$ (similarly for $V_{\uparrow_\pi}$, exchanging π-maximizing for π-minimizing).

Optimal value functions in BMDPs

We now consider how to define an optimal value function for a BMDP. First, consider the expression $\max_{\pi \in \Pi}(V_{\updownarrow_\pi})$. This expression is ill-formed because we have not defined how to rank the interval value functions $V_{\updownarrow_\pi}$ in order to select a maximum.[4] We focus

[4] Similar issues arise if we attempt to define the optimal value function using a Bellman style equation such as Eq. (3) because we must compute a maximization over a set of intervals.

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:

$$\begin{aligned}&\big([l_1,u_1]\leqslant_{\text{pes}}[l_2,u_2]\big)\Leftrightarrow\big(l_1<l_2\text{ or }(l_1=l_2\text{ and }u_1\leqslant u_2)\big),\\&\big([l_1,u_1]\leqslant_{\text{opt}}[l_2,u_2]\big)\Leftrightarrow\big(u_1<u_2\text{ or }(u_1=u_2\text{ and }l_1\leqslant l_2)\big).\end{aligned}\tag{17}$$

We extend these orderings to partial orders over interval value functions by relating two value functions $V_{\updownarrow 1}\leqslant_{\text{opt}}V_{\updownarrow 2}$ only when $V_{\updownarrow 1}(q)\leqslant_{\text{opt}}V_{\updownarrow 2}(q)$ for every state q. We can now use either of these orderings to compute $\max_{\pi\in\Pi}(V_{\updownarrow\pi})$, yielding two definitions of optimal value function and optimal policy. However, since the orderings are partial (on value functions), we prove first (Theorem 8) that the set of policies contains a policy that achieves the desired maximum under each ordering (i.e., a policy whose interval value function is ordered above that of every other policy).

Definition 6. An *optimistically optimal policy* π_{opt} is any policy such that $V_{\updownarrow\pi_{\text{opt}}}\geqslant_{\text{opt}}V_{\updownarrow\pi}$ for all policies π. A *pessimistically optimal policy* π_{pes} is any policy such that $V_{\updownarrow\pi_{\text{pes}}}\geqslant_{\text{pes}}V_{\updownarrow\pi}$ for all policies π.

In Theorem 8, we prove that there exist optimistically optimal policies by induction (an analogous proof holds for pessimistically optimal policies). We develop this proof in two stages, mirroring the two-stage definition of $\geqslant_{\text{opt}}$ (first emphasizing the upper bound and then breaking ties with the lower bound). We first construct a policy π' for which the upper bounds of the interval value function $V_{\uparrow\pi'}$ dominate those $V_{\uparrow\pi''}$ of any other policy π''. We then show that the finite set of such policies (all tied on upper bounds) can be combined to construct a policy π_{opt} with the same upper bound values $V_{\uparrow\pi_{\text{opt}}}$ and whose lower bounds $V_{\downarrow\pi_{\text{opt}}}$ dominate those of any other policy. Each of these constructions relies on the following policy composition operator:

Definition 7. Let $\oplus_{\text{opt}}$ and $\oplus_{\text{pes}}$ denote composition operators on policies, defined as follows. Consider policies $\pi_1,\pi_2\in\Pi$.

Let $\pi_3=\pi_1\oplus_{\text{opt}}\pi_2$ if for all states $p\in Q$:

$$\pi_3(p)=\begin{cases}\pi_1(p) & \text{if } V_{\updownarrow\pi_1}(p)\geqslant_{\text{opt}}V_{\updownarrow\pi_2}(p),\\ \pi_2(p) & \text{otherwise.}\end{cases}\tag{18}$$

Let $\pi_3=\pi_1\oplus_{\text{pes}}\pi_2$ if for all states $p\in Q$:

$$\pi_3(p)=\begin{cases}\pi_1(p) & \text{if } V_{\updownarrow\pi_1}(p)\geqslant_{\text{pes}}V_{\updownarrow\pi_2}(p),\\ \pi_2(p) & \text{otherwise.}\end{cases}\tag{19}$$

Our task would be relatively easy if it were necessarily true that

$$V_{\updownarrow(\pi_1\oplus_{\text{opt}}\pi_2)}\geqslant_{\text{opt}}V_{\updownarrow\pi_1}\quad\text{and}\quad V_{\updownarrow(\pi_1\oplus_{\text{opt}}\pi_2)}\geqslant_{\text{opt}}V_{\updownarrow\pi_2}\tag{20}$$

(and likewise for the pessimistic case). However, because of the lexicographic nature of $\geqslant_{\text{opt}}$, these statements do not hold (in particular, the lower bound values for some states may be worse in the composed policy than in either component even when the upper bounds on those states do not change). For this reason, we prove a somewhat weaker result that must be used in a two-stage fashion as demonstrated below:

Lemma 3. *Given a BMDP $M_{\updownarrow}$, and policies $\pi_1, \pi_2 \in \Pi$, $\pi_3 = \pi_1 \oplus_{\text{opt}} \pi_2$, and $\pi_4 = \pi_1 \oplus_{\text{pes}} \pi_2$:*

(a) $V_{\uparrow\pi_3} \geqslant_{\text{dom}} V_{\uparrow\pi_1}$ *and* $V_{\uparrow\pi_3} \geqslant_{\text{dom}} V_{\uparrow\pi_2}$.
(b) *If* $V_{\uparrow\pi_1} = V_{\uparrow\pi_2}$ *then* $V_{\updownarrow\pi_3} \geqslant_{\text{opt}} V_{\updownarrow\pi_1}$ *and* $V_{\updownarrow\pi_3} \geqslant_{\text{opt}} V_{\updownarrow\pi_2}$.
(c) $V_{\downarrow\pi_4} \geqslant_{\text{dom}} V_{\downarrow\pi_1}$ *and* $V_{\downarrow\pi_4} \geqslant_{\text{dom}} V_{\downarrow\pi_2}$.
(d) *If* $V_{\downarrow\pi_1} = V_{\downarrow\pi_2}$ *then* $V_{\updownarrow\pi_4} \geqslant_{\text{pes}} V_{\updownarrow\pi_1}$ *and* $V_{\updownarrow\pi_4} \geqslant_{\text{pes}} V_{\updownarrow\pi_2}$.

Proof. See Appendix A. □

Theorem 8. *There exists at least one optimistically* (*pessimistically*) *optimal policy.*

Proof. Enumerate Π as a finite sequence of policies $\pi_1, \ldots, \pi_k$. Consider composing these policies together to construct the policy $\pi_{\text{opt,up}}$ as follows:

$$\pi_{\text{opt,up}} = \big(\big((\pi_1 \oplus_{\text{opt}} \pi_2) \oplus_{\text{opt}} \cdots\big) \oplus_{\text{opt}} \pi_k\big). \tag{21}$$

Note that $\pi_{\text{opt,up}}$ may depend on the ordering of $\pi_1, \ldots, \pi_k$, but that any ordering is satisfactory for this proof. It is straightforward to show by induction using Lemma 3 that $V_{\uparrow\pi_{\text{opt,up}}} \geqslant_{\text{dom}} V_{\uparrow\pi_i}$ for each $1 \leqslant i \leqslant k$. Now enumerate the subset of Π for which the value function upper bounds equal those of $\pi_{\text{opt,up}}$, i.e., enumerate $\{\pi' \mid V_{\uparrow\pi'} = V_{\uparrow\pi_{\text{opt,up}}}\}$ as $\{\pi'_1, \ldots, \pi'_l\}$. Consider again composing the policies π'_i together as above to form the policy π_{opt}:

$$\pi_{\text{opt}} = \big(\big((\pi'_1 \oplus_{\text{opt}} \pi'_2) \oplus_{\text{opt}} \cdots\big) \oplus_{\text{opt}} \pi'_l\big). \tag{22}$$

It is again straightforward to show using Lemma 3 that $V_{\downarrow\pi_{\text{opt}}} \geqslant_{\text{dom}} V_{\downarrow\pi'_i}$ for each $1 \leqslant i \leqslant l$. It follows immediately that $V_{\updownarrow\pi_{\text{opt}}} \geqslant_{\text{opt}} V_{\updownarrow\pi}$ for every $\pi \in \Pi$, as desired. A similar construction using $\oplus_{\text{pes}}$ yields a pessimistically optimal policy π_{pes}. □

Theorem 8 justifies the following definition:

Definition 8. The *optimistic optimal value function* $V_{\updownarrow\text{opt}}$ and the *pessimistic optimal value function* $V_{\updownarrow\text{pes}}$ are given by:

$$V_{\updownarrow\text{opt}} = \max_{\pi\in\Pi}(V_{\updownarrow\pi}) \quad \text{using } \leqslant_{\text{opt}} \text{ to order interval value functions;}$$

$$V_{\updownarrow\text{pes}} = \max_{\pi\in\Pi}(V_{\updownarrow\pi}) \quad \text{using } \leqslant_{\text{pes}} \text{ to order interval value functions.}$$

The above two notions of optimal value can be understood in terms of a two-player game in which the first player chooses a policy π and then the second player chooses the MDP M in $M_{\updownarrow}$ in which to evaluate the policy π (see Shapley's work [16] for the

origins of this viewpoint). The goal for the first player is to get the highest[5] resulting value function $V_{M,\pi}$. The upper bounds $V_{\uparrow \mathrm{opt}}$ of the optimistically optimal value function represent the best value function the first player can obtain in this game if the second player cooperates by selecting an MDP to maximize $V_{M,\pi}$ (the lower bound $V_{\downarrow \mathrm{opt}}$ corresponds to how badly this optimistic strategy for the first player can misfire if the second player betrays the first player and selects an MDP to minimize $V_{M,\pi}$). The lower bounds $V_{\downarrow \mathrm{pes}}$ of the pessimistically optimal value function represent the best the first player can do under the assumption that the second player is an adversary, trying to minimize the resulting value function.

We conclude this section by stating a Bellman equation theorem for the optimal interval value functions just defined. The equations below form the basis for our iterative algorithm for computing the optimal interval value functions for a BMDP. We start by stating two definitions that are useful in proving the Bellman theorem as well as in later sections. It is useful to have notation to denote the set of actions that maximize the upper bound at each state. For a given value function V, we write ρ_V for the function from states to sets of actions such that for each state p,

$$\rho_V(p) = \operatorname*{argmax}_{\alpha \in A} \max_{M \in M_\updownarrow} VI_{M,\alpha}(V)(p). \tag{23}$$

Likewise, for the pessimistic case, we define σ_V for the function from states to sets of actions giving the actions that maximize the lower bound. For each state p, $\sigma_V(p)$ is given by

$$\sigma_V(p) = \operatorname*{argmax}_{\alpha \in A} \min_{M \in M_\updownarrow} VI_{M,\alpha}(V)(p). \tag{24}$$

Theorem 9. *For any BMDP $M_\updownarrow$, the following Bellman-like equations hold at every state p,*

$$V_{\updownarrow \mathrm{opt}}(p) = \max_{\alpha \in A,\, \leqslant_{\mathrm{opt}}} \Big[\min_{M \in M_\updownarrow} VI_{M,\alpha}(V_{\downarrow \mathrm{opt}})(p),\ \max_{M \in M_\updownarrow} VI_{M,\alpha}(V_{\uparrow \mathrm{opt}})(p) \Big], \tag{25}$$

and

$$V_{\updownarrow \mathrm{pes}}(p) = \max_{\alpha \in A,\, \leqslant_{\mathrm{pes}}} \Big[\min_{M \in M_\updownarrow} VI_{M,\alpha}(V_{\downarrow \mathrm{pes}})(p),\ \max_{M \in M_\updownarrow} VI_{M,\alpha}(V_{\uparrow \mathrm{pes}})(p) \Big]. \tag{26}$$

Proof. See Appendix A. □

5. Estimating interval value functions

In this section, we describe dynamic programming algorithms that operate on bounded-parameter MDPs. We first define the interval equivalent of policy evaluation $IVI_{\updownarrow \pi}$ which computes $V_{\updownarrow \pi}$, and then define the variants $IVI_{\updownarrow \mathrm{opt}}$ and $IVI_{\updownarrow \mathrm{pes}}$ which compute the optimistic and pessimistic optimal value functions.

[5] Value functions are ranked by $\geqslant_{\mathrm{dom}}$.

5.1. Interval policy evaluation

In direct analogy to the exact MDP definition of VI_π in Section 3, we define a function $IVI_{\updownarrow\pi}$ (for *interval value iteration*) which maps interval value functions to other interval value functions. We prove that iterating $IVI_{\updownarrow\pi}$ on any initial interval value function produces a sequence of interval value functions that converges to $V_{\updownarrow\pi}$ in a polynomial number of steps, given a fixed discount factor γ.

$IVI_{\updownarrow\pi}(V_\updownarrow)$ is an interval value function, defined for each state p as follows:

$$IVI_{\updownarrow\pi}(V_\updownarrow)(p) = \Big[\min_{M\in M_\updownarrow} VI_{M,\pi}(V_\downarrow)(p),\ \max_{M\in M_\updownarrow} VI_{M,\pi}(V_\uparrow)(p)\Big]. \tag{27}$$

We define $IVI_{\downarrow\pi}$ and $IVI_{\uparrow\pi}$ to be the corresponding mappings from value functions to value functions (note that for input $V_\updownarrow$, $IVI_{\downarrow\pi}$ does not depend on $V_\uparrow$ and so can be viewed as a function from $\overline{V}$ to $\overline{V}$—likewise for $IVI_{\uparrow\pi}$ and $V_\downarrow$).

The algorithm to compute $IVI_{\updownarrow\pi}$ 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 $M_\updownarrow$ that minimizes (maximizes) the value attained. We select such an MDP by selecting a transition probability function F within the bounds specified by the $F_\updownarrow$ component of $M_\updownarrow$ to minimize (maximize) the value—each possible way of selecting F corresponds to one MDP in $M_\updownarrow$. We can select the values of $F_{pq}(\alpha)$ 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 fixed α and p, the value of $F_{pq}(\alpha)$ for each state q so as to minimize (maximize) the expression $\sum_{q\in Q}(F_{pq}(\alpha)V(q))$. This step constitutes the heart of the $IVI_{\updownarrow\pi}$ algorithm and the only significant way the algorithm differs from standard policy evaluation by successive approximation by iterating $VI_{M,\pi}$.

To compute the lower bounds $IVI_{\downarrow\pi}$ the idea is to sort the possible destination states q into increasing order according to their $V_\downarrow$ value, and then choose the transition probabilities within the intervals specified by $F_\updownarrow$ so as to send as much probability mass to the states early in the ordering as possible (upper bounds are computed similarly, but sorting the states into decreasing order by their $V_\uparrow$ value). Let $O = q_1, q_2, \ldots, q_k$ be such an ordering of Q—so that for all i and j if $1 \leqslant i \leqslant j \leqslant k$ then $V_\downarrow(q_i) \leqslant V_\downarrow(q_j)$ (increasing order). We can then show that the order-maximizing MDP M_O is the MDP that minimizes the desired expression $\sum_{q\in Q}(F^M_{pq}(\alpha)V(q))$. The order-maximizing MDP for the decreasing order based on $V_\uparrow$ will maximize the same expression to generate the upper bound in Eq. (27).

Fig. 6 illustrates the basic iterative step in the above algorithm, for the upper bound, i.e., maximizing, case. The states q_i are ordered according to the value estimates in $V_\uparrow$. 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. For a more precise account of the algorithm, please refer to Fig. 7 for a pseudocode description of the computation of $IVI_{\updownarrow\pi}(V_\updownarrow)$.

Techniques similar to those in Section 3 can be used to prove that iterating $IVI_{\downarrow\pi}$ (or $IVI_{\uparrow\pi}$) converges to $V_{\downarrow\pi}$ (or $V_{\uparrow\pi}$). The key theorems, stated below, assert first that $IVI_{\downarrow\pi}$ is a contraction mapping, and second that $V_{\downarrow\pi}$ is a fixed point of $IVI_{\downarrow\pi}$ and are easily proven.

Theorem 10. *For any policy π, $IVI_{\downarrow\pi}$ and $IVI_{\uparrow\pi}$ are contraction mappings.*

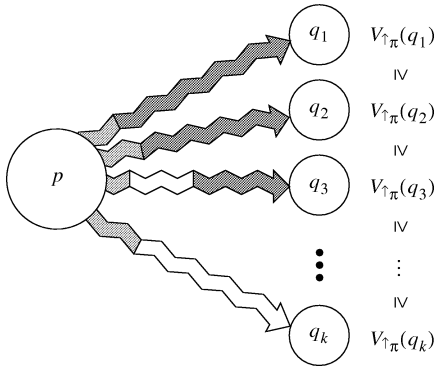

Fig. 6. An illustration of the basic dynamic programming step in computing an approximate value function for a fixed policy and bounded-parameter MDP. $V_{\uparrow\pi}$ gives the upper bounds of the current interval estimates of V_π. The lighter shaded portions of each arc represent the required lower 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.

Proof. See Appendix A. □

Theorem 11. *For any policy π, $V_{\downarrow\pi}$ is a fixed point of $IVI_{\downarrow\pi}$ and $V_{\uparrow\pi}$ of $IVI_{\uparrow\pi}$, and therefore $V_{\updownarrow\pi}$ is a fixed point of $IVI_{\updownarrow\pi}$.*

Proof. See Appendix A. □

These theorems, together with Theorem 1 (the Banach fixed point theorem) imply that iterating $IVI_{\updownarrow\pi}$ on any initial interval value function converges to $V_{\updownarrow\pi}$, regardless of the starting point.

Theorem 12. *For fixed $\gamma < 1$, interval policy evaluation converges to the desired interval 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 BMDP parameters.*

Proof (*Sketch*). We provide only the key ideas behind this proof.

(a) By Theorem 10, $IVI_{\updownarrow\pi}$ is a contraction by γ on both the upper and lower bound value functions, and thus the successive estimates of $V_{\updownarrow\pi}$ produced converge exponentially to the unique fixed point.
(b) By Theorem 11, the unique fixed point is the desired value function.
(c) The upper bound and lower bound value functions making up the true $V_{\updownarrow\pi}$ are the value functions of π in particular MDPs (π-maximizing and π-minimizing MDPs, respectively) in $X_{M_\updownarrow}$.
(d) The parameters for the MDPs in $X_{M_\updownarrow}$ can be specified with a number of bits polynomial in the number of bits used to specify the BMDP parameters.
(e) The value function for a policy in an MDP can be written as the solution to a linear program. The precision of any such solution can be bounded in terms of the number

```
IVI↕(V↕, π)
\\we assume that V↕ is represented as:
\\   V↓ is a vector of n real numbers giving lower bounds for states q1 to qn
\\   V↑ is a vector of n real numbers giving upper bounds for states q1 to qn
{   Create O, a vector of n states for holding a permutation of the states q1 to qn

    \\first, compute new lower bounds
    O = sort_increasing_order(q1, ..., qn, <lb);    \\ <lb compares state lower bounds
    Update(V↓, π, O);

    \\second, compute new upper bounds
    O = sort_decreasing_order(q1, ..., qn, <ub);    \\ <ub compares state upper bounds
    Update(V↑, π, O)}
```

```
\\ Update(v, π, o) updates v using the order-maximizing MDP for o
\\   o is a state ordering—a vector of states (a permutation of q1, ..., qn)
\\   v is a value function—a vector of real numbers of length n

Update(v, π, o)
{   Create F', a matrix of n by n real numbers

    \\ the next loop sets F' to describe π in the order-maximizing MDP for o
    for each state p {
       used = Σ_state q F↓ p,q(π(p));
       remaining = 1 − used;

       \\ distribute remaining probability mass to states early in the ordering
       for i = 1 to n {          \\ i is used to index into ordering o
          min = F↓ p,o(i)(π(p));
          desired = F↑ p,o(i)(π(p)) ;
          if (desired ⩽ remaining)
             then F'(p, o(i)) = min + desired;
             else F'(p, o(i)) = min + remaining;
          remaining = max(0, remaining − desired)}}

    \\ F' now describes π in the order-maximizing MDP with respect to O,
    \\ finally, update v using a value iteration-like update based on F'
    for each state p
       v(p) = R(p) + γ Σ_state q F'(p, q)v(q)    }
```

Fig. 7. Pseudocode for one iteration of interval policy evaluation ($IVI_{\updownarrow}$).

of bits used to specify the linear program. This precision bound allows the definition of a stopping condition for $IVI_{\updownarrow\pi}$ when adequate precision is obtained. □

5.2. Interval value iteration

As in the case of altering VI_π to obtain VI, it is straightforward to modify $IVI_{\updownarrow\pi}$ 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.

$$IVI_{\updownarrow_{\text{opt}}}(V_{\updownarrow})(p) = \max_{\alpha\in A,\, \leqslant_{\text{opt}}} \Big[\min_{M\in M_{\updownarrow}} VI_{M,\alpha}(V_{\downarrow})(p),\ \max_{M\in M_{\updownarrow}} VI_{M,\alpha}(V_{\uparrow})(p)\Big], \tag{28}$$

$$IVI_{\updownarrow_{\text{pes}}}(V_{\updownarrow})(p) = \max_{\alpha\in A,\, \leqslant_{\text{pes}}} \Big[\min_{M\in M_{\updownarrow}} VI_{M,\alpha}(V_{\downarrow})(p),\ \max_{M\in M_{\updownarrow}} VI_{M,\alpha}(V_{\uparrow})(p)\Big]. \tag{29}$$

The added maximization step introduces no new difficulties in implementing the algorithm —for more details we provide pseudocode for $IVI_{\updownarrow_{\text{opt}}}$ in Fig. 8. We discuss convergence for $IVI_{\updownarrow_{\text{opt}}}$—the convergence results for $IVI_{\updownarrow_{\text{pes}}}$ are similar. We first summarize our approach and then cover the same ground in more detail.

We write $IVI_{\uparrow_{\text{opt}}}$ for the upper bound returned by $IVI_{\updownarrow_{\text{opt}}}$, and we consider $IVI_{\uparrow_{\text{opt}}}$ a function from $\overline{V}$ to $\overline{V}$ because $IVI_{\uparrow_{\text{opt}}}(V_{\updownarrow})$ depends only on $V_{\uparrow}$ due to the way $\leqslant_{\text{opt}}$ compares intervals primarily based on their upper bound. $IVI_{\uparrow_{\text{opt}}}$ can easily be shown to be a contraction mapping, and it can be shown that $V_{\uparrow_{\text{opt}}}$ is a fixed point of $IVI_{\uparrow_{\text{opt}}}$. It then follows that $IVI_{\uparrow_{\text{opt}}}$ converges to $V_{\uparrow_{\text{opt}}}$ (and we can argue as for $IVI_{\updownarrow_{\pi}}$ that this convergence occurs in polynomially many steps for fixed γ). The analogous results for $IVI_{\downarrow_{\text{opt}}}$ are somewhat more problematic. Because the action selection is done according to $\leqslant_{\text{opt}}$, which focuses primarily on the interval upper bounds, $IVI_{\downarrow_{\text{opt}}}$ is not properly a mapping from $\overline{V}$ to $\overline{V}$, as the action choice for $IVI_{\downarrow_{\text{opt}}}(V_{\updownarrow})$ depends on both $V_{\downarrow}$ and $V_{\uparrow}$. In particular, for each state, the action that maximizes the lower bound is chosen from among the subset of actions that (equally) maximize the upper bound.

To deal with this complication, we observe that if we fix the upper bound value function $V_{\uparrow}$, we can view $IVI_{\downarrow_{\text{opt}}}$ as a function from $\overline{V}$ to $\overline{V}$ carrying the lower bounds of the input value function to the lower bounds of the output. To formalize this idea, we introduce some new notation. First, given two value functions V_1 and V_2 we define the interval value function $[V_1, V_2]$ to be the function from states p to intervals $[V_1(p), V_2(p)]$ (this notation is essentially the inverse of the $\downarrow$ and $\uparrow$ notation which extracts lower and upper bound functions from interval functions). Using this new notation, we define a family $\{IVI_{\downarrow_{\text{opt},V}}\}$ of functions from $\overline{V}$ to $\overline{V}$, indexed by a value function V. For each value function V, we define $IVI_{\downarrow_{\text{opt},V}}(V')$ to be the function from $\overline{V}$ to $\overline{V}$ that maps V' to $IVI_{\downarrow_{\text{opt}}}([V', V])$. (Analogously, we define $IVI_{\uparrow_{\text{pes},V}}(V')$ to map V' to $IVI_{\uparrow_{\text{pes}}}([V, V'])$.) We note that $IVI_{\downarrow_{\text{opt},V}}$ has the following relationships to $IVI_{\updownarrow_{\text{opt}}}$:

$$\begin{aligned} IVI_{\updownarrow_{\text{opt}}}(V_{\updownarrow}) &= \big[IVI_{\downarrow_{\text{opt},V_{\uparrow}}}(V_{\downarrow}), IVI_{\uparrow_{\text{opt}}}(V_{\uparrow})\big], \\ IVI_{\downarrow_{\text{opt}}}(V_{\updownarrow}) &= IVI_{\downarrow_{\text{opt},V_{\uparrow}}}(V_{\downarrow}). \end{aligned} \tag{30}$$

In analyzing $IVI_{\updownarrow_{\text{opt}}}$, we also use the notation defined in Section 4 for the set of actions that maximize the upper bound at each state. We restate the relevant definition here for convenience. For a given value function V, we write ρ_V for the function from states to sets of actions such that for each state p,

$$\rho_V(p) = \operatorname*{argmax}_{\alpha\in A} \max_{M\in M_{\updownarrow}} VI_{M,\alpha}(V)(p). \tag{31}$$

Likewise, for the pessimistic case, we defined σ_V in Section 4.

```
IVI↕opt(V↕)
\\we assume that V↕ is represented as:
\\   V↓ is a vector of n real numbers giving lower bounds for states q1 to qn
\\   V↑ is a vector of n real numbers giving upper bounds for states q1 to qn
{   Create O, a vector of n states for holding a permutation of the states q1 to qn

    \\first, compute new lower bounds
    O = sort_increasing_order(q1, ..., qn, <lb);     \\ <lb compares state lower bounds
    VI-Update(V↓, O)};

    \\second, compute new upper bounds
    O = sort_decreasing_order(q1, ..., qn, <ub);     \\ <ub compares state upper bounds
    VI-Update(V↑, O)}
```

```
\\ VI-Update(v, o) updates v using the order-maximizing MDP for o
\\   o is a state ordering—a vector of states (a permutation of q1, ..., qn)
\\   v is a value function—a vector of real numbers of length n

VI-Update(v, o)
{   Create Fa, a matrix of n by n real numbers for each action a

    \\the next loop sets each Fa to describe a in the order-maximizing MDP for o
    for each state p and action a {
       used = Σstate q F↓p,q(a);
       remaining = 1 − used;

       \\ distribute remaining probability mass to states earlier in ordering
       for i = 1 to n {          \\ i is used to index into ordering o
          min = F↓p,o(i)(a);
          desired = F↑p,o(i)(a);
          if (desired ⩽ remaining)
             then Fa(p, o(i)) = min + desired;
             else Fa(p, o(i)) = min + remaining;
          remaining = max(0, remaining − desired)}}

    \\ Fa now describes a in the order-maximizing MDP with respect to O,
    \\ finally, update v using a value iteration-like update based on F′
    for each state p
       v(p) = max_{a∈A} [ R(p) + γ Σ_{state q} Fa(p, q)v(q)   } ]
```

Fig. 8. Pseudocode: an iteration of optimistic interval value iteration ($IVI_{\updownarrow \text{opt}}$).

Given the definition of $\leqslant_{\text{opt}}$, it is straightforward to show the following lemma.

Lemma 4. *For any value functions V, V' and state p,*

$$\begin{aligned} IVI_{\downarrow \text{opt},V}(V')(p) &= \max_{\alpha \in \rho_V(p)} \min_{M \in M_\updownarrow} VI_{M,\alpha}(V')(p), \\ IVI_{\uparrow \text{pes},V}(V')(p) &= \max_{\alpha \in \sigma_V(p)} \min_{M \in M_\updownarrow} VI_{M,\alpha}(V')(p). \end{aligned} \tag{32}$$

Proof. By inspection of the definitions of $IVI_{\updownarrow_{\text{opt}}}$ and $IVI_{\updownarrow_{\text{pes}}}$. □

We now show that for each V, $IVI_{\downarrow_{\text{opt},V}}$ is a contraction mapping relative to the sup norm, and thus converges to a unique fixed point, as desired. Theorem 9 then implies that $V_{\updownarrow_{\text{opt}}}$ is the unique fixed point found. ($V_{\updownarrow_{\text{pes}}}$ in the case of $IVI_{\updownarrow_{\text{pes}}}$). We then show that at any point after polynomially many iterations of $IVI_{\updownarrow_{\text{opt}}}$, the resulting interval value function $V_{\updownarrow}$ has upper bounds $V_{\uparrow}$ that have converged to a fixed point of $IVI_{\uparrow_{\text{opt}}}$, and thus further iteration of $IVI_{\updownarrow_{\text{opt}}}$ is equivalent to iterating $IVI_{\uparrow_{\text{opt}}}$ and $IVI_{\downarrow_{\text{opt},V_{\uparrow}}}$ together in parallel to generate the upper and lower bounds, respectively. We can also show that for any V, polynomially many iterations of $IVI_{\downarrow_{\text{opt},V}}$ suffice for convergence to a fixed point. Similar results hold for $IVI_{\updownarrow_{\text{pes}}}$. We now give the details of these results.

Theorem 13.
(a) *$IVI_{\uparrow_{\text{opt}}}$ and $IVI_{\downarrow_{\text{pes}}}$ are contraction mappings.*
(b) *For any value function V and associated action set selection function ρ_V and σ_V, $IVI_{\downarrow_{\text{opt},V}}$ and $IVI_{\uparrow_{\text{pes},V}}$ are contraction mappings.*

Proof. See Appendix A. □

Theorem 14. *For fixed γ, polynomially many iterations of $IVI_{\updownarrow_{\text{opt}}}$ can be used to find $V_{\updownarrow_{\text{opt}}}$, and polynomially many iterations of $IVI_{\updownarrow_{\text{pes}}}$ can be used to find $V_{\updownarrow_{\text{pes}}}$, with both polynomials defined relative to the problem size including the number of bits used in specifying the parameters.*

Proof (*Sketch*). The argument here is exactly as in Theorem 12, relying on Theorems 9 and 13, except that the iterations must be taken to convergence in two stages. Considering $IVI_{\updownarrow_{\text{opt}}}$, we must first iterate until the upper bound has converged, with the polynomial-time bound on iterations deriving by a similar argument to the proof of Theorem 12; then once the upper bounds have converged we must then iterate until the lower bounds have converged, again in polynomially many iterations by another argument similar to that in the proof of Theorem 12.

More precisely, let $V_{\updownarrow_1}, V_{\updownarrow_2}, \ldots$, be a sequence of interval value functions found by iterating $IVI_{\updownarrow_{\text{opt}}}$, so that for each i greater or equal to 1 we have $V_{\updownarrow_{i+1}}$ equal to $IVI_{\updownarrow_{\text{opt}}}(V_{\updownarrow_i})$. Then an argument similar to the proof of Theorem 12 guarantees that for some j polynomial in the size of the problem, $V_{\updownarrow_j}$ must have upper bounds that are equal to the true fixed point upper bound values, up to the maximum precision of the true fixed point. We then know that truncating the upper value bounds in $V_{\updownarrow_j}$ to that precision (to get an interval value function $V'_{\updownarrow_1}$) gives the true fixed point upper bound values. We can then iterate $IVI_{\updownarrow_{\text{opt}}}$ starting on $V'_{\updownarrow_1}$ to get another sequence of value functions where the upper bounds are unchanging and the lower bounds are converging to the correct fixed point values in the same manner.

A similar argument shows polynomial convergence for $IVI_{\updownarrow_{\text{pes}}}$. □

6. Policy selection

In this section, we consider the problem of selecting a policy based on the value bounds computed by our IVI algorithms. This section is not intended as an additional research contribution as much as a discussion of issues that arise in solving BMDP problems and of alternative approaches to policy selection (other than the optimistic and pessimistic approaches we take here). 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. We note again that in many applications it is possible to compute prior probabilities over these parameters, but that these computations are prohibitively expensive in our motivating application (solving large state-space problems by approximate state-space aggregation).

Despite the fact that a BMDP does not specify which particular MDP we are facing, 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 ultimately have to carry out the 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 maximize potential performance, we might assume that the outside process cooperates by choosing the MDP in order to help us; we can then select the policy that performs as well as possible given that assumption. In contrast, we might minimize the risk of performing poorly by thinking in adversarial terms: we can select the policy that performs as well as possible under the assumption that an adversary chooses the MDP so that we perform as poorly as possible (in each case we assume that the MDP is chosen from the BMDP family of MDPs *after* the policy has been selected in order to minimize/maximize the value of that policy).

These choices correspond to optimistic and pessimistic optimal policies as defined above. 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 straight-forward manner to extract policies that achieve those values.

We note that it may seem unnatural to be required to take an optimistic or a pessimistic approach in order to select a policy—certainly this is not analogous to policy selection for standard MDPs. This requirement grows out of our model assumption that we have no prior probabilities on the model parameters, and we have argued that this assumption is in fact natural at very least in our motivating domain of approximate state-space aggregation. The same assumption is also natural in performing sensitivity analysis, as described in the next section. We also note that there is precedent in the related MDP literature for considering optimistic and pessimistic approaches to policy selection in the face of uncertainty about the model; see, for example, the work of Satia and Lave in [15].

Alternative approaches to selecting a policy are possible, but some approaches that seem natural at first run into trouble. For instance, we might consider placing a uniform prior probability on each model parameter within its specified interval. Unfortunately, the model parameters cannot in general be selected independently (because they must together

represent a well-formed probability distribution after selection), and there may not even be any joint prior distribution over the parameters which marginalizes to the uniform distribution over the provided intervals when marginalized to each parameter. Therefore, the uniform distribution over the provided intervals does not enjoy any distinguished status—it may not even correspond to a well-formed prior over the underlying MDPs in the BMDP family.

There are other well-formed choices corresponding to other means of totally ordering real closed intervals (other than $\leqslant_{\text{opt}}$ and $\leqslant_{\text{pes}}$). For instance, we might order intervals by their midpoints, asserting a preference for states where the highest and lowest value possible in the underlying MDP family have a high mean. It is not clear when this choice might be prefered; however, we believe our methods can be naturally adapted to compute optimal policy values for other interval orderings, if desired.

A natural goal would be to find 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. Moreover, it is not at all clear how to find such a policy—our methods do not appear to generalize in this direction. As noted just above, this goal does *not* correspond to assuming a uniform prior over the model parameters, but rather a more complex joint distribution over the parameters. Also, this average case solution would not in general provide useful information in our motivating application of state-space aggregation: we would have no guarantee that the uniform prior over MDP models consistent with the BMDP had any useful correlation with the original large MDP that aggregated to the BMDP. In contrast, as discussed below, the optimistic and pessimistic bounds we compute apply directly to any MDP when the BMDP analyzed is formed by state-space aggregation of that MDP. Nevertheless, the question of how to compute the optimal average case policy for a BMDP appears to be a useful direction for future research.

7. Prototype implementation results and potential applications

In this section we discuss our intended applications for the new BMDP algorithms, and present empirical results from a prototype implementation of the algorithms for use in state-space aggregation. We note that no particular difficulties were encountered in implementing the new BMDP algorithms—implementation is more demanding than that of standard MDP algorithms, but only by the addition of a sorting algorithm.

Sensitivity analysis

One way in which bounded-parameter MDPs might be useful in planning under uncertainty might 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 assess this possible sensitivity to the parameters, we might perturb the MDP parameters and evaluate the policy with respect to the perturbed MDP. Alternatively, we could use BMDPs 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. We note that this use provides our original motivation for developing BMDPs, and therefore it is this use that we give prototype empirical results for below.

In the state-aggregate interpretation of a BMDP, 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 $M_{\updownarrow}$ 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 $F_{\updownarrow B_i B_j}(\alpha)$ represent the interval value for the transition from B_i to B_j on action α defined as follows:

$$F_{\updownarrow B_i B_j}(\alpha) = \left[\min_{p \in B_i} \sum_{q \in B_j} F_{pq}(\alpha), \max_{p \in B_i} \sum_{q \in B_j} F_{pq}(\alpha) \right]. \tag{33}$$

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.

In Dean et al. [10] we discuss methods for using an implicit representation of a exact MDP with a large number of states to construct an explicit BMDP with a possibly much smaller number of states based on an aggregation method. We then show that policies computed for this BMDP can be extended to the original 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 the state spaces). In particular, the lower interval bounds computed on a given state block by $IVI_{\downarrow \text{pes}}$ give lower bounds on the optimal value for states in that block in the original MDP; likewise, the upper interval bounds computed by $IVI_{\uparrow \text{opt}}$ give upper bounds on the optimal value in the original MDP.

Empirical results

We constructed a prototype implementation of our BMDP algorithms, interval value iteration and interval policy evaluation. We then used this implementation in conjunction with implementations of our previously presented approximate state-space aggregation algorithms [10] in order to compute lower and upper bounds on the values of individual states in large MDP problems.

Table 1
Model size after approximate minimization

# State Vars	# States	$\varepsilon = 0$	$\varepsilon = 0.01$	$\varepsilon = 0.1$	$\varepsilon = 0.3$	$\varepsilon = 0.5$	$\varepsilon = 0.8$
9	512	114	114	72	24	11	8
10	1024	131	122	85	55	21	21
13	8192	347	347	272	148	66	63
14	16384			442	153	67	63
15	32768			520	152	88	69
IVI Inaccuracy:		0%	0.2%	10%	40%	58%	62%

The MDP problems used were derived by partially modelling air campaign planning problems using implicit MDP representations. These problems involve selecting tasks for a variety of military aircraft over time in order to maximize the utility of their actions, and require modeling many aspects of the aircraft capabilities, resources, crew, and tasks. Modeling the full problem as an MDP is still out of reach—the MDP models used in these experiments were constructed by representing the problem at varying degrees of (extremely coarse) abstraction so that the resulting problem would be within reach of our prototype implementation.

We show in Table 1 the original problem state-space size, the state-space size of the BMDP that results from our aggregation algorithm, and the quality of the resulting state-value bounds for several different sized MDP problems. Each row in the table corresponds to a specific explicit MDP that we solved (approximately and/or exactly) using state-space aggregation. We note that one parameter (ε) of our aggregation method is the degree of approximation tolerated in transition probability—this corresponds to the interval width in the BMDP parameter intervals. As this parameter is given larger and larger values across the columns of the table, the aggregate BMDP model has fewer and fewer states—in return, the value bounds obtained are less and less tight. The quality of the resulting state-value bounds is given by showing "IVI Inaccuracy"—this percentage is the average width of the value intervals computed as a percentage of the difference between the lowest possible state value and the highest possible state value (these are defined by assuming a repeated occurence of the lowest/highest reward available for an infinite time period and computing the total discounted reward obtained). Our prototype aggregation code was incapable of handling the exact and near-exact analysis of the largest models tried, and those entries in the table are therefore missing.

We note that IVI inaccuracies of much greater than 25% may not represent very useful bounds on state value (we have not yet conducted experiments to evaluate this question). For this reason, the last three columns of the table are shown primarily for completeness and to satisfy curiosity. However, an inaccuracy of 10% can be expected to yield useful information in selecting between different control actions—we can think of this level of inaccuracy as allowing us to rate each state on a scale of one to ten as to how good its value is. Such ratings should be very useful in designing control policies.

We note that our prototype code is not optimized in its handling of either space or time. Similar prototype code for explicit MDP problems can handle no more than a few hundred

states. Production versions of explicit MDP code today can handle as many as a million or so states. Our aggregation and BMDP algorithms, even in this unoptimized form, are able to obtain nontrivial bounds on state value for state-space sizes involving thousands of states. We believe that a production version of these algorithms could derive near-optimal policies for MDP planning problems involving hundreds of millions of states.

8. 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 of the closest such ideas. First, BMDPs specialize the MDPs with imprecisely known parameters (MDPIPs) described and analyzed in the operations research literature by White and Eldeib [17,18], and Satia and Lave [15]. The more general MDPIPs described in these papers require more general and expensive algorithms for solution. For example, [17] allows an arbitrary linear program to define the bounds on the transition probabilities (and allows no imprecision in the reward parameters)—as a result, the solution technique presented appeals to linear programming at each iteration of the solution algorithm rather than exploit the specific structure available in a BMDP as we do here. [15] mentions the restriction to BMDPs but gives no special algorithms to exploit this restriction. Their general MDPIP algorithm is very different from our algorithm and involves two nested phases of policy iteration—the outer phase selecting a traditional policy and the inner phase selecting a "policy" for "nature", i.e., a choice of the transition parameters to minimize or maximize value (depending on whether optimistic or pessimistic assumptions prevail). Our work, while originally developed independently of the MDPIP literature, follows similar lines to [15] in defining optimistic and pessimistic optimal policies. In summary, when uncertainty about MDP parameters is such that a BMDP model is appropriate, the MDPIP literature does not provide an approach that exploits the restricted structure to achieve an efficient method (we note appealing to linear programming at each iteration can be very expensive).

Shapley [16] introduced the notion of *stochastic games* to describe two-person games in which the transition probabilities are controlled by the two players. MDPIPs, and therefore BMDPs, are a special case of *alternating* stochastic games in which the first player is the decision-making agent and the second player, often considered as either an adversary or advocate, makes its move by choosing from the set of possible MDPs consistent with having seen the agent's move.

Bertsekas and Castañon [3] 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 [13] describes their use in solving partially observable MDPs. Puterman [14] provides an excellent introduction to Markov decision processes and techniques involving bounding value functions.

Boutilier, Dean and Hanks [5] provide a careful treatment of MDP-related methods demonstrating how they provide a unifying framework for modeling a wide range of problems in AI involving planning under uncertainty. This paper also describes such related

issues as state space aggregation, decomposition and abstraction as these ideas pertain to work in AI. We encourage the reader unfamiliar with the connection between classical planning methods in AI and Markov decision processes to refer to this paper.

Boutilier and Dearden [6] and Boutilier et al. [8] describe methods for solving implicitly described MDPs using dynamic aggregation—in their methods the state space aggregates vary over the iterations of the dynamic programming algorithm. This work can be viewed as using a compact representation of both policies and value functions in terms of state aggregates to perform the familiar dynamic programming algorithms. Dean and Givan [9] reinterpret this work in terms of computing explicitly described MDPs with aggregate states corresponding to the aggregates that the above compactly represented value functions use when they have converged. Dean, Givan, and Leach [10] discuss relaxing these aggregation techniques to construct approximate aggregations—it is from this work that the notion of BMDP emerged in order to represent the resulting aggregate models.

Bounded-parameter MDPs allow us to represent uncertainty about or variation in the parameters of a Markov decision process. Interval value functions capture the resulting variation in policy values. In this paper, we have defined both bounded-parameter MDP and interval value function, and given algorithms for computing interval value functions, and selecting and evaluating policies.

Acknowledgements

Many thanks to Michael Littman for useful conversation and insights concerning the proofs that the algorithms herein run in polynomial time.

Appendix A. Proofs omitted above for readability

Lemma 1. *For any policy $\pi \in \Pi$, MDP $M \in M_{\updownarrow}$, and value function $v \in \overline{V}$,*

(a) *there are MDPs $M_1 \in X_{M_{\updownarrow}}$ and $M_2 \in X_{M_{\updownarrow}}$ such that*

$$V_{M_1,\pi} \leqslant_{\mathrm{dom}} V_{M,\pi} \leqslant_{\mathrm{dom}} V_{M_2,\pi}. \tag{10}$$

(b) *Also, there are MDPs $M_3 \in X_{M_{\updownarrow}}$ and $M_4 \in X_{M_{\updownarrow}}$ such that*

$$VI_{M_3,\pi}(v) \leqslant_{\mathrm{dom}} VI_{M,\pi}(v) \leqslant_{\mathrm{dom}} VI_{M_4,\pi}(v). \tag{11}$$

Proof. To show the existence of M_1, let $O = q_1, \ldots, q_k$ be an ordering on states such that for all i and j if $1 \leqslant i \leqslant j \leqslant k$ then $V_{M,\pi}(q_i) \leqslant V_{M,\pi}(q_j)$ (increasing order). Note that ties in state values permit different orderings; for the proof, it is sufficient to chose one ordering arbitrarily. Consider $M_O \in X_{M_{\updownarrow}}$, the order-maximizing MDP of O. M_O is constructed so as to send as much probability mass as possible to states earlier in the ordering O, i.e., to those states q with lower value $V_{M,\pi}(q)$. It follows that for any state p,

$$\sum_{q\in Q} \left(F_{pq}^{M_O}\left(\pi(p)\right)V_{M,\pi}(q)\right) \leqslant \sum_{q\in Q} \left(F_{pq}^{M}\left(\pi(p)\right)V_{M,\pi}(q)\right). \tag{A.1}$$

Thus, for any state p,

$$V_{M,\pi}(p) = R(p) + \gamma \sum_{q \in Q} \left(F^{M}_{pq}\left(\pi(p)\right) V_{M,\pi}(q)\right) \tag{A.2}$$

$$\geqslant R(p) + \gamma \sum_{q \in Q} \left(F^{M_O}_{pq}\left(\pi(p)\right) V_{M,\pi}(q)\right) \tag{A.3}$$

$$= VI_{M_O,\pi}(V_{M,\pi})(p). \tag{A.4}$$

By Theorem 6, these lines imply $V_{M_O,\pi} \leqslant_{\text{dom}} V_{M,\pi}$, as desired.

The existence of M_2 can be shown in the same except that O is chosen to order the states by increasing value. Thus M_O is constructed so that

$$\sum_{q \in Q} \left(F^{M}_{pq}\left(\pi(p)\right) V_{M,\pi}(q)\right) \leqslant \sum_{q \in Q} \left(F^{M_O}_{pq}\left(\pi(p)\right) V_{M,\pi}(q)\right). \tag{A.5}$$

Part (b) is shown in the same manner as part (a) except that we replace each occurrence of $V_{M,\pi}(p)$ with $VI_{M,\pi}(v)(p)$ and each occurrence of $V_{M,\pi}(q)$ with $v(q)$. □

Lemma 2. *Let π be a policy in Π and M_1, M_2 be MDPs in $M_{\updownarrow}$.*

(a) *For $M_3 = M_1 \oplus^{\pi}_{\max} M_2$,*

$$V_{M_3,\pi} \geqslant_{\text{dom}} V_{M_1,\pi} \quad \textit{and} \quad V_{M_3,\pi} \geqslant_{\text{dom}} V_{M_2,\pi}, \quad \textit{and} \tag{14}$$

(b) *for $M_3 = M_1 \oplus^{\pi}_{\min} M_2$,*

$$V_{M_3,\pi} \leqslant_{\text{dom}} V_{M_1,\pi} \quad \textit{and} \quad V_{M_3,\pi} \leqslant_{\text{dom}} V_{M_2,\pi}. \tag{15}$$

Proof. (a) We construct a value function v such that $v \geqslant_{\text{dom}} V_{M_1,\pi}$, $v \geqslant_{\text{dom}} V_{M_2,\pi}$, and $v \leqslant_{\text{dom}} V_{M_3,\pi}$, as follows. For each $p \in Q$, let

$$v(p) = \max\left(V_{M_1,\pi}(p), V_{M_2,\pi}(p)\right). \tag{A.6}$$

Note that this implies $v \geqslant_{\text{dom}} V_{M_1,\pi}$ and $v \geqslant_{\text{dom}} V_{M_2,\pi}$. We now show using Theorem 6 that $v \leqslant_{\text{dom}} V_{M_3,\pi}$. By Theorem 6 it suffices to prove that $v \leqslant_{\text{dom}} VI_{M_3,\pi}(v)$, which we now do by showing $v(p) \leqslant VI_{M_3,\pi}(v)(p)$ for arbitrary $p \in Q$.

Case 1: We suppose $V_{M_1,\pi}(p) \geqslant V_{M_2,\pi}(p)$.

From Eq. (A.6) we then have that $v(p) = V_{M_1,\pi}(p)$. By the definition of $\oplus^{\pi}_{\max}$, we know $F^{M_3}_{pq}(\pi(p)) = F^{M_1}_{pq}(\pi(p))$ when $V_{M_1,\pi}(p) \geqslant V_{M_2,\pi}(p)$ as in this case. This fact, together with the definitions of VI, $V_{M_1,\pi}$, $\oplus^{\pi}_{\max}$, and v allow the following chain of equations to conclude the proof of case 1:

$$\begin{aligned} v(p) &= V_{M_1,\pi}(p) \\ &= R(p) + \gamma \sum_{q \in Q} F^{M_1}_{pq}\left(\pi(p)\right) V_{M_1,\pi}(q) \\ &\leqslant R(p) + \gamma \sum_{q \in Q} F^{M_1}_{pq}\left(\pi(p)\right) v(q) \\ &= R(p) + \gamma \sum_{q \in Q} F^{M_3}_{pq}\left(\pi(p)\right) v(q) \\ &= VI_{M_3,\pi}(v)(p). \end{aligned} \tag{A.7}$$

Case 2: Suppose $V_{M_1,\pi}(p) < V_{M_2,\pi}(p)$.

We then have $F_{pq}^{M_3}(\pi(p)) = F_{pq}^{M_2}(\pi(p))$ by the definition of $\oplus_{\max}^{\pi}$, and $v(p) = V_{M_1,\pi}$ by the definition of v, and Eq. (A.7) holds with M_1 replaced by M_2, as desired, concluding the proof of part (a).

(b) The proof is exactly dual to part (a) by replacing "max" with "min, $\leqslant$ with $\geqslant$ (and vice versa), and $<$ with $>$. □

Lemma 3. *Given a BMDP $M_{\updownarrow}$, and policies $\pi_1, \pi_2 \in \Pi$, $\pi_3 = \pi_1 \oplus_{\text{opt}} \pi_2$, and $\pi_4 = \pi_1 \oplus_{\text{pes}} \pi_2$,*

(a) $V_{\uparrow\pi_3} \geqslant_{\text{dom}} V_{\uparrow\pi_1}$ *and* $V_{\uparrow\pi_3} \geqslant_{\text{dom}} V_{\uparrow\pi_2}$.
(b) *If* $V_{\uparrow\pi_1} = V_{\uparrow\pi_2}$ *then* $V_{\updownarrow\pi_3} \geqslant_{\text{opt}} V_{\updownarrow\pi_1}$ *and* $V_{\updownarrow\pi_3} \geqslant_{\text{opt}} V_{\updownarrow\pi_2}$.
(c) $V_{\downarrow\pi_4} \geqslant_{\text{dom}} V_{\downarrow\pi_1}$ *and* $V_{\downarrow\pi_4} \geqslant_{\text{dom}} V_{\downarrow\pi_2}$.
(d) *If* $V_{\downarrow\pi_1} = V_{\downarrow\pi_2}$ *then* $V_{\updownarrow\pi_4} \geqslant_{\text{pes}} V_{\updownarrow\pi_1}$ *and* $V_{\updownarrow\pi_4} \geqslant_{\text{pes}} V_{\updownarrow\pi_2}$.

Proof. (a) We prove part (a) of the lemma by constructing a value function v such that $v \geqslant_{\text{dom}} V_{\uparrow\pi_1}$ and $v \geqslant_{\text{dom}} V_{\uparrow\pi_2}$. We then show that $v \leqslant_{\text{dom}} V_{\uparrow\pi_3}$ using Theorem 6. We construct v as follows. Let

$$v(p) = \max\bigl(V_{\uparrow\pi_1}(p), V_{\uparrow\pi_2}(p)\bigr) \quad \text{for each } p \in Q.$$

This construction implies that $v \geqslant_{\text{dom}} V_{\uparrow\pi_1}$ and $v \geqslant_{\text{dom}} V_{\uparrow\pi_2}$. We now show $v \leqslant_{\text{dom}} V_{\uparrow\pi_3}$ by giving an MDP M_3 for which $V_{M_3,\pi_3} \geqslant_{\text{dom}} v$. Using Theorem 6 it suffices to show that $VI_{M_3,\pi_3}(v) \geqslant_{\text{dom}} v$.

Let $M_1 \in M_{\updownarrow}$ be a π_1-maximizing MDP, and $M_2 \in M_{\updownarrow}$ be a π_2-maximizing MDP. Note that this implies that $V_{\uparrow\pi_1} = V_{M_1,\pi_1}$ and $V_{\uparrow\pi_2} = V_{M_2,\pi_2}$.

We now construct $M_3 \in M_{\updownarrow}$ as follows: for each p, q, α,

$$F_{pq}^{M_3}(\alpha) = \begin{cases} F_{pq}^{M_1}(\alpha) & \text{if } V_{\updownarrow\pi_1}(p) \geqslant_{\text{opt}} V_{\updownarrow\pi_2}(p), \\ F_{pq}^{M_2}(\alpha) & \text{otherwise.} \end{cases}$$

It remains to show that $VI_{M_3,\pi_3}(v)(p) \geqslant v(p)$ for all $p \in Q$. Now fix an arbitrary $p \in Q$.

Case 1: Suppose $V_{\updownarrow\pi_1}(p) \geqslant_{\text{opt}} V_{\updownarrow\pi_2}(p)$.

Then by the definition of $\oplus_{\text{opt}}$, $\pi_3(p) = \pi_1(p)$. Also, by the definition of $\geqslant_{\text{opt}}$, $V_{\uparrow\pi_1}(p) \geqslant V_{\uparrow\pi_2}(p)$, and so $v(p) = V_{M_1,\pi_1}(p)$ is true, and by the definition of M_3, $F_{pq}^{M_3}(\pi_3(p)) = F_{pq}^{M_1}(\pi_3(p))$. The following inequations thus hold:

$$\begin{aligned} v(p) &= V_{\uparrow\pi_1}(p) && \text{(A.8)} \\ &= R(p) + \gamma \sum_{q\in Q} \bigl(F_{pq}^{M_1}\bigl(\pi_1(p)\bigr) V_{\uparrow\pi_1}(q)\bigr) && \text{(A.9)} \\ &= R(p) + \gamma \sum_{q\in Q} \bigl(F_{pq}^{M_3}\bigl(\pi_3(p)\bigr) V_{\uparrow\pi_1}(q)\bigr) && \text{(A.10)} \\ &\leqslant R(p) + \gamma \sum_{q\in Q} \bigl(F_{pq}^{M_3}\bigl(\pi_3(p)\bigr) v(q)\bigr) && \text{(A.11)} \\ &= VI_{M_3,\pi_3}(v)(p). && \text{(A.12)} \end{aligned}$$

Case 2: Suppose $V_{\updownarrow\pi_1}(p) <_{\text{opt}} V_{\updownarrow\pi_2}(p)$.

Then by the definition of $\oplus_{\text{opt}}$, $\pi_3(p) = \pi_2(p)$. Also, by the definition of $\geqslant_{\text{opt}}$, $V_{\uparrow\pi_1}(p) \leqslant V_{\uparrow\pi_2}(p)$, and so $v(p) = V_{M_2,\pi_2}(p)$ is true, and by the definition of M_3, $F_{pq}^{M_3}(\pi_3(p)) = F_{pq}^{M_2}(\pi_3(p))$. Then Eqs. (A.8)–(A.12) hold with M_2 and π_2 in place of M_1 and π_1 respectively, yielding again that $v(p) \leqslant VI_{M_3,\pi_3}(v)(p)$, as desired.

Case 1 and Case 2 together imply that $v(p) \leqslant VI_{M_3,\pi_3}(v)(p)$ for all $p \in Q$, which with Theorem 6 implies part (a) of the lemma.

(b) Supposing that $V_{\uparrow\pi_1} = V_{\uparrow\pi_2}$, we show $V_{\updownarrow\pi_3} \geqslant_{\text{opt}} V_{\updownarrow\pi_1}$ and $V_{\updownarrow\pi_3} \geqslant_{\text{opt}} V_{\updownarrow\pi_2}$. From part (a) of the theorem, we know that $V_{\uparrow\pi_3} \geqslant_{\text{dom}} V_{\uparrow\pi_1}$ and $V_{\uparrow\pi_3} \geqslant_{\text{dom}} V_{\uparrow\pi_2}$. It suffices to prove in addition that $V_{\downarrow\pi_3} \geqslant_{\text{dom}} V_{\downarrow\pi_1}$ and $V_{\downarrow\pi_3} \geqslant_{\text{dom}} V_{\downarrow\pi_2}$. We show both by defining $v(p) = \max(V_{\downarrow\pi_1}(p), V_{\downarrow\pi_2}(p))$ for each state $p \in Q$, observing that $v \geqslant_{\text{dom}} V_{\downarrow\pi_1}$ and $v \geqslant_{\text{dom}} V_{\downarrow\pi_2}$, and then showing that $V_{\downarrow\pi_3} \geqslant_{\text{dom}} v$.

We can show $V_{\downarrow\pi_3} \geqslant_{\text{dom}} v$ by showing that for arbitrary $M \in M_{\updownarrow}$, $V_{M,\pi_3} \geqslant_{\text{dom}} v$. By Theorem 6 it suffices to show that for arbitrary state $p \in Q$, $VI_{M,\pi_3}(v)(p) \geqslant v$. We divide now into two cases:

Case 1: Suppose $V_{\downarrow\pi_1}(p) \geqslant V_{\downarrow\pi_2}(p)$.

With the part (b) assumption ($V_{\uparrow\pi_1} = V_{\uparrow\pi_2}$), this implies $V_{\updownarrow\pi_1}(p) \geqslant_{\text{opt}} V_{\updownarrow\pi_2}(p)$. Then by the definition of $\oplus_{\text{opt}}$, $\pi_3(p) = \pi_1(p)$. Also by definition in this case $v(p) = V_{\downarrow\pi_1}(p)$. Let M_1 be a π_1-minimizing MDP. The following inequation chain gives the desired conclusion:

$$v(p) = V_{\downarrow\pi_1}(p) \tag{A.13}$$

$$= R(p) + \gamma \sum_{q\in Q} F_{pq}^{M_1}\big(\pi_1(p)\big) V_{\downarrow\pi_1}(q) \tag{A.14}$$

$$\leqslant R(p) + \gamma \sum_{q\in Q} F_{pq}^{M}\big(\pi_1(p)\big) V_{\downarrow\pi_1}(q) \tag{A.15}$$

$$\leqslant R(p) + \gamma \sum_{q\in Q} F_{pq}^{M}\big(\pi_3(p)\big) v(q) \tag{A.16}$$

$$= VI_{M,\pi_3}(v)(p). \tag{A.17}$$

Line (A.15) requires some justification. Consider an MDP M_1' defined to agree with M_1 everywhere except that $F_{pq}^{M_1'} = F_{pq}^{M}$ for every $q \in Q$. If line (A.15) did not hold, we would have $VI_{M_1',\pi_1}(V_{\downarrow\pi_1}) <_{\text{dom}} V_{\downarrow\pi_1}$ and then Theorem 6 could be used to show that $V_{M_1',\pi_1} <_{\text{dom}} V_{\downarrow\pi_1}$, contradicting the definition of $V_{\downarrow\pi_1}$.

Case 2: Suppose $V_{\downarrow\pi_1}(p) < V_{\downarrow\pi_2}(p)$.

With the part (b) assumption this implies that $V_{\updownarrow\pi_1}(p) <_{\text{opt}} V_{\updownarrow\pi_2}(p)$.

Then by the definition of $\oplus_{\text{opt}}$, $\pi_3(p) = \pi_2(p)$. Also $v(p) = V_{\downarrow\pi_2}(p)$. Let M_2 be a π_2-minimizing MDP. Eqs. (A.13)–(A.17) now hold with M_1 and π_1 replaced by M_2 and π_2, respectively.

We have now shown in both cases that $v(p) \leqslant VI_{M,\pi_3}(v)(p)$, as desired, concluding the proof of part (b) of the theorem.

(c) We prove part (c) of the lemma by constructing a value function v such that $v \geqslant_{\text{dom}} V_{\downarrow\pi_1}$ and $v \geqslant_{\text{dom}} V_{\downarrow\pi_2}$. We then show that $v \leqslant_{\text{dom}} V_{\downarrow\pi_4}$ using Theorem 6. We

construct v as follows. Let $v(p) = \max(V_{\downarrow\pi_1}(p), V_{\downarrow\pi_2}(p))$ for each $p \in Q$. This implies $v \geqslant_{\text{dom}} V_{\downarrow\pi_1}$ and $v \geqslant_{\text{dom}} V_{\downarrow\pi_2}$. We now show $v \leqslant_{\text{dom}} V_{\downarrow\pi_4}$ by showing that for arbitrary $M \in M_{\updownarrow}$, $V_{M,\pi_4} \geqslant_{\text{dom}} v$. Using Theorem 6 it suffices to show that $VI_{M,\pi_4}(v) \geqslant_{\text{dom}} v$.

Let $M_1 \in M_{\updownarrow}$ be a π_1-minimizing MDP, and $M_2 \in M_{\updownarrow}$ be a π_2-minimizing MDP. Note that this implies that $V_{\downarrow\pi_1} = V_{M_1,\pi_1}$ and $V_{\downarrow\pi_2} = V_{M_2,\pi_2}$.

Now fix an arbitrary $p \in Q$, and show that $VI_{M,\pi_4}(v)(p) \geqslant v(p)$.

Case 1: Suppose $V_{\updownarrow\pi_1}(p) \geqslant_{\text{pes}} V_{\updownarrow\pi_2}(p)$.

Then by the definition of $\oplus_{\text{pes}}$, $\pi_4(p) = \pi_1(p)$. Also, by the definition of $\geqslant_{\text{pes}}$, $V_{\downarrow\pi_1}(p) \geqslant V_{\downarrow\pi_2}(p)$, and so $v(p) = V_{M_1,\pi_1}(p)$ is true. Eqs. (A.13)–(A.17) now hold with π_4 in place of π_3, giving the desired result.

Case 2: Suppose $V_{\updownarrow\pi_1}(p) <_{\text{pes}} V_{\updownarrow\pi_2}(p)$.

Then by the definition of $\oplus_{\text{pes}}$, $\pi_4(p) = \pi_2(p)$. Also, by the definition of $\geqslant_{\text{pes}}$, $V_{\downarrow\pi_1}(p) \leqslant V_{\downarrow\pi_2}(p)$, and so $v(p) = V_{M_2,\pi_2}(p)$ is true. Then Eqs. (A.13)–(A.17) hold with M_2, π_2, and π_4 in place of M_1, π_1, and π_3, respectively, yielding again that $v(p) \leqslant VI_{M,\pi_4}(v)(p)$, as desired.

Case 1 and Case 2 together imply that $v(p) \leqslant VI_{M,\pi_4}(v)(p)$ for all $p \in Q$, which with Theorem 6 implies part (c) of the theorem.

(d) Supposing that $V_{\downarrow\pi_1} = V_{\downarrow\pi_2}$, we show $V_{\updownarrow\pi_4} \geqslant_{\text{pes}} V_{\updownarrow\pi_1}$ and $V_{\updownarrow\pi_4} \geqslant_{\text{pes}} V_{\updownarrow\pi_2}$. From part (c) of the theorem, we know that $V_{\downarrow\pi_4} \geqslant_{\text{dom}} V_{\downarrow\pi_1}$ and $V_{\downarrow\pi_4} \geqslant_{\text{dom}} V_{\downarrow\pi_2}$. It suffices to prove in addition that $V_{\uparrow\pi_4} \geqslant_{\text{dom}} V_{\uparrow\pi_1}$ and $V_{\uparrow\pi_4} \geqslant_{\text{dom}} V_{\uparrow\pi_2}$. We show both by defining $v(p) = \max(V_{\uparrow\pi_1}(p), V_{\uparrow\pi_2}(p))$ for each state $p \in Q$, observing that $v \geqslant_{\text{dom}} V_{\uparrow\pi_1}$ and $v \geqslant_{\text{dom}} V_{\uparrow\pi_1}$, and then showing that $V_{\uparrow\pi_4} \geqslant_{\text{dom}} v$ by giving an MDP M_4 for which $V_{M_4,\pi_4} \geqslant_{\text{dom}} v$. Using Theorem 6 it suffices to show that $VI_{M_4,\pi_4}(v) \geqslant_{\text{dom}} v$.

Let $M_1 \in M_{\updownarrow}$ be a π_1-maximizing MDP, and $M_2 \in M_{\updownarrow}$ be a π_2-maximizing MDP. Note that this implies that $V_{\uparrow\pi_1} = V_{M_1,\pi_1}$ and $V_{\uparrow\pi_2} = V_{M_2,\pi_2}$.

We now construct $M_4 \in M_{\updownarrow}$ as follows: for each p, q, α,

$$F_{pq}^{M_4}(\alpha) = \begin{cases} F_{pq}^{M_1}(\alpha) & \text{if } V_{\updownarrow\pi_1}(p) \geqslant_{\text{pes}} V_{\updownarrow\pi_2}(p), \\ F_{pq}^{M_2}(\alpha) & \text{otherwise.} \end{cases}$$

It remains to show that $VI_{M_4,\pi_4}(v)(p) \geqslant v(p)$ for all $p \in Q$. Now fix an arbitrary $p \in Q$.

Case 1: Suppose $V_{\uparrow\pi_1}(p) \geqslant_{\text{pes}} V_{\uparrow\pi_2}(p)$.

With the part (d) assumption this implies that $V_{\updownarrow\pi_1}(p) \geqslant_{\text{pes}} V_{\updownarrow\pi_2}(p)$. Then by the definition of $\oplus_{\text{pes}}$, $\pi_4(p) = \pi_1(p)$. Also by definition in this case $v(p) = V_{\uparrow\pi_1}(p)$. Also, by the definition of M_4, $F_{pq}^{M_4}(\pi_4(p)) = F_{pq}^{M_1}(\pi_4(p))$. Eqs. (A.8)–(A.12) with π_3 and M_3 replaced by π_4 and M_4 complete the argument.

Case 2: Suppose $V_{\uparrow\pi_1}(p) < V_{\uparrow\pi_2}(p)$.

With the part (d) assumption this implies that $V_{\updownarrow\pi_1}(p) <_{\text{pes}} V_{\updownarrow\pi_2}(p)$.

Then by definition $\pi_4(p) = \pi_2(p)$. Also $v(p) = V_{\uparrow\pi_2}(p)$. Eqs. (A.8)–(A.12) now hold with M_1, $M_3\pi_1$, and π_3 replaced by M_2, $M_4\pi_2$, and π_4, respectively.

We have now shown in both cases that $v(p) \leqslant VI_{M_4,\pi_4}(v)(p)$, as desired, concluding the proof of part (d) of the theorem. □

Theorem 9. *For any BMDP $M_{\updownarrow}$, at every state p,*

$$V_{\updownarrow_{\mathrm{opt}}}(p) = \max_{\alpha\in A,\ \leqslant_{\mathrm{opt}}} \Big[\min_{M\in M_{\updownarrow}} VI_{M,\alpha}(V_{\downarrow_{\mathrm{opt}}})(p),\ \max_{M\in M_{\updownarrow}} VI_{M,\alpha}(V_{\uparrow_{\mathrm{opt}}})(p)\Big], \tag{25}$$

and

$$V_{\updownarrow_{\mathrm{pes}}}(p) = \max_{\alpha\in A,\ \leqslant_{\mathrm{pes}}} \Big[\min_{M\in M_{\updownarrow}} VI_{M,\alpha}(V_{\downarrow_{\mathrm{pes}}})(p),\ \max_{M\in M_{\updownarrow}} VI_{M,\alpha}(V_{\uparrow_{\mathrm{pes}}})(p)\Big]. \tag{26}$$

Proof. We consider the $V_{\updownarrow_{\mathrm{opt}}}$ version only. Throughout this proof we assume π_{opt} is an optimistically optimal policy for $M_{\updownarrow}$, which exists by Theorem 8. We suppose Eq. (25) is false and show a contradiction. We have two cases:

Case 1: Suppose the upper bounds are not equal at some state p:

$$V_{\uparrow_{\mathrm{opt}}}(p) \neq \max_{\alpha\in A}\max_{M\in M_{\updownarrow}} VI_{M,\alpha}(V_{\uparrow_{\mathrm{opt}}})(p). \tag{A.18}$$

There are two ways this can happen:

Subcase 1(a): Suppose there exist some MDP $M \in M_{\updownarrow}$ and action $\alpha \in A$ such that

$$V_{\uparrow_{\mathrm{opt}}}(p) < VI_{M,\alpha}(V_{\uparrow_{\mathrm{opt}}})(p). \tag{A.19}$$

We show how to construct a policy π whose interval value $V_{\updownarrow_\pi}$ dominates $V_{\updownarrow_{\mathrm{opt}}}$ under $\leqslant_{\mathrm{opt}}$, contradicting the definition of $V_{\updownarrow_{\mathrm{opt}}}$. Define π to be the same as π_{opt} except that $\pi(p) = \alpha$. By the definition of $V_{\updownarrow_{\pi_{\mathrm{opt}}}}$, there must exist $M' \in M_{\updownarrow}$ such that $V_{\uparrow_{\mathrm{opt}}} = V_{\uparrow_{\pi_{\mathrm{opt}}}} = V_{M',\pi_{\mathrm{opt}}}$. From the theory of exact MDPs, we then have that:

$$V_{\uparrow_{\mathrm{opt}}} = V_{M',\pi_{\mathrm{opt}}} = VI_{M',\pi_{\mathrm{opt}}}(V_{M',\pi_{\mathrm{opt}}}) = VI_{M',\pi_{\mathrm{opt}}}(V_{\uparrow_{\mathrm{opt}}}). \tag{A.20}$$

Our subcase assumption implies

$$V_{\uparrow_{\mathrm{opt}}}(p) < VI_{M,\pi}(V_{\uparrow_{\mathrm{opt}}})(p). \tag{A.21}$$

Consider the MDP $M_3 \in M_{\updownarrow}$ with the same parameters as M' except at state p where the parameters are given by M. More formally,

$$F^{M_3}_{p'q'} = \begin{cases} F^{M}_{p'q'} & \text{when } p' = p, \\ F^{M}_{p'q'} & \text{otherwise.} \end{cases} \tag{A.22}$$

This construction of M_3, together with Eqs. (A.20) and (A.21), guarantees the following property of $V_{\uparrow_{\mathrm{opt}}}$:

$$V_{\uparrow_{\mathrm{opt}}} <_{\mathrm{dom}} VI_{M_3,\pi}(V_{\uparrow_{\mathrm{opt}}}). \tag{A.23}$$

Eq. (A.23) along with Theorem 6 implies that $V_{M_3,\pi} >_{\mathrm{dom}} V_{\uparrow_{\mathrm{opt}}}$ and thus that $V_{\updownarrow_\pi} >_{\mathrm{opt}} V_{\updownarrow_{\mathrm{opt}}}$, contradicting the definition of $V_{\updownarrow_{\mathrm{opt}}}$ and concluding Subcase 1(a).

Subcase 1(b). Suppose that for every choice of $\alpha \in A$ and $M \in M_{\updownarrow}$

$$V_{\uparrow_{\mathrm{opt}}}(p) > VI_{M,\alpha}(V_{\uparrow_{\mathrm{opt}}})(p). \tag{A.24}$$

We obtain a contradiction directly by exhibiting α and $M \in M_{\updownarrow}$ in violation of this supposition. Let α be $\pi_{\mathrm{opt}}(p)$. Let M be a π_{opt}-maximizing MDP in $M_{\updownarrow}$, which exists by Theorem 7. Our selection of π_{opt} guarantees that $V_{\uparrow_{\pi_{\mathrm{opt}}}} = V_{\uparrow_{\mathrm{opt}}}$, and our choice of M

guarantees that $V_{M,\pi_{\text{opt}}} = V_{\uparrow \pi_{\text{opt}}}$. Eqs. (7) and (8) from the theory of exact MDPs then ensure that $V_{\uparrow \text{opt}}(p) = VI_{M,\alpha}(V_{\uparrow \text{opt}})(p)$, concluding Case 1.

Case 2. Suppose at every state q the upper bounds are equal but at some state p the lower bounds are not equal:

$$\begin{aligned} \text{for all } q, \quad & V_{\uparrow \text{opt}}(q) = \max_{\alpha \in A} \max_{M \in M_{\updownarrow}} VI_{M,\alpha}(V_{\uparrow \text{opt}})(q), \quad \text{and} \\ & V_{\downarrow \text{opt}}(p) \neq \max_{\alpha \in \rho_{V_{\uparrow \text{opt}}}(p)} \min_{M \in M_{\updownarrow}} VI_{M,\alpha}(V_{\downarrow \text{opt}})(p). \end{aligned} \tag{A.25}$$

Note that the action selection in the second line of Eq. (A.25) is restricted to range over those actions in $\rho_{V_{\uparrow \text{opt}}}(p)$ because those are the only actions that can be selected in Eq. (25) due to the emphasis of $\leqslant_{\text{opt}}$ on upper bounds (the upper bounds achievable by an action primarily determine whether it is selected by the outer maximization in Eq. (25), and only if the action is tied for the maximum upper bound, i.e., in $\rho_{V_{\uparrow \text{opt}}}(p)$, does its lower bound affect the maximization).

Again, there are two ways the second line of Eq. (A.25) can hold.

Subcase 2(a). Suppose $V_{\downarrow \text{opt}}(p)$ is too small, i.e., there exists some action $\alpha \in \rho_{V_{\uparrow \text{opt}}}(p)$ such that for every MDP $M \in M_{\updownarrow}$, we have

$$V_{\downarrow \text{opt}}(p) < VI_{M,\alpha}(V_{\downarrow \text{opt}})(p). \tag{A.26}$$

We show a contradiction by giving a policy π whose interval value function is greater than $V_{\updownarrow \text{opt}}$ under the $\leqslant_{\text{opt}}$ ordering. Define π to be the same as π_{opt} except that $\pi(p) = \alpha$. By the definition of $V_{\updownarrow \pi_{\text{opt}}}$, there must exist $M' \in M_{\updownarrow}$ such that $V_{\uparrow \text{opt}} = V_{\uparrow \pi_{\text{opt}}} = V_{M',\pi_{\text{opt}}}$. As in Subcase 1(a), we then have that:

$$V_{\uparrow \text{opt}} = V_{M',\pi_{\text{opt}}} = VI_{M',\pi_{\text{opt}}}(V_{M',\pi_{\text{opt}}}) = VI_{M',\pi_{\text{opt}}}(V_{\uparrow \text{opt}}). \tag{A.27}$$

From Eq. (A.25) and $\alpha \in \rho_{V_{\uparrow \text{opt}}}(p)$ it follows that for some $M \in M_{\updownarrow}$,

$$V_{\uparrow \text{opt}}(p) = VI_{M,\alpha}(V_{\uparrow \text{opt}})(p), \tag{A.28}$$

and thus for $M_3 \in M_{\updownarrow}$ defined as in Subcase 1(a) to be equal to M' everywhere except at state p where M_3 is equal to M, we have

$$V_{\uparrow \text{opt}} = VI_{M_3,\pi}(V_{\uparrow \text{opt}}). \tag{A.29}$$

Therefore $V_{M_3,\pi} = V_{\uparrow \text{opt}}$, and by the definitions of $V_{\updownarrow \text{opt}}$ and $V_{\uparrow}\pi$, we then have that $V_{\uparrow \text{opt}} \geqslant_{\text{dom}} V_{\uparrow \pi} \geqslant_{\text{dom}} V_{M_3,\pi} = V_{\uparrow \text{opt}}$, and so $V_{\uparrow \pi}$ is equal to $V_{\uparrow \text{opt}}$. We must now show that $V_{\downarrow \pi} >_{\text{dom}} V_{\downarrow \text{opt}}$ to conclude Subcase 2(a). We show this by showing that for every MDP $M_4 \in M_{\updownarrow}$, $V_{\downarrow \text{opt}} <_{\text{dom}} VI_{M_4,\pi}(V_{\downarrow \text{opt}})$ and using Theorem 6 to conclude $V_{M_4,\pi} >_{\text{dom}} V_{\downarrow \text{opt}}$ and thus $V_{\downarrow \pi} >_{\text{dom}} V_{\downarrow \text{opt}}$ as desired.

To conclude Subcase 2(a), then, we must show $V_{\downarrow \text{opt}} <_{\text{dom}} VI_{M_4,\pi}(V_{\downarrow \text{opt}})$. We show this by contradiction. Suppose this is false—then either $V_{\downarrow \text{opt}} = VI_{M_4,\pi}(V_{\downarrow \text{opt}})$, which our Subcase 2(a) assumption rules out at state p, or there must be some state q for which $V_{\downarrow \text{opt}}(q) > VI_{M_4,\pi}(V_{\downarrow \text{opt}})(q)$. Again our subcase assumption rules this out for state p, so we know that q is not equal to p, and therefore by our choice of π we have that $\pi(q) = \pi_{\text{opt}}(q)$, and thus that $V_{\downarrow \text{opt}}(q) > VI_{M_4,\pi_{\text{opt}}}(V_{\downarrow \text{opt}})(q)$. We can now derive a contradiction

by combining M_4 at state q with a π_{opt}-minimizing MDP M_5 at all other states to get an MDP $M_6 \in M_{\updownarrow}$ for which $V_{\downarrow \text{opt}}$ strictly dominates $VI_{M_6,\pi_{\text{opt}}}(V_{\downarrow \text{opt}})$, showing that $V_{\downarrow \text{opt}} >_{\text{dom}} V_{M_6,\pi_{\text{opt}}}$ (by Theorem 6) contradicting the fact that $V_{\downarrow \pi_{\text{opt}}} = V_{\downarrow \text{opt}}$. (The combination of M_4 and M_5 to get M_6 is analogous to the construction in line (A.22) above.)

Subcase 2(b). Suppose $V_{\downarrow \text{opt}}(p)$ is "too big" in line (A.25), i.e., for every action $\alpha \in \rho_{V_{\uparrow \text{opt}}}(p)$ there is some MDP $M_\alpha \in M_{\updownarrow}$ such that $VI_{M_\alpha,\alpha}(V_{\downarrow \text{opt}})(p) < V_{\downarrow \text{opt}}(p)$. Consider $\alpha = \pi_{\text{opt}}(p)$. The definition of "optimistically optimal" along with the theory of exact MDPs guarantees us that there is some MDP M such that

$$V_{\uparrow \text{opt}} = V_{\uparrow \pi_{\text{opt}}} = V_{M,\pi_{\text{opt}}} = VI_{M,\pi_{\text{opt}}}(V_{M,\pi_{\text{opt}}}) = VI_{M,\pi_{\text{opt}}}(V_{\uparrow \text{opt}}). \tag{A.30}$$

By our Case 2 assumption,

$$V_{\uparrow \text{opt}}(p) = \max_{\alpha \in A} \max_{M \in M_{\updownarrow}} VI_{M,\alpha}(V_{\uparrow \text{opt}})(p), \tag{A.31}$$

and this, together with line (A.30) and $\alpha = \pi_{\text{opt}}(p)$ implies

$$VI_{M,\pi_{\text{opt}}}(V_{\uparrow \text{opt}})(p) = \max_{\alpha \in A} \max_{M \in M_{\updownarrow}} VI_{M,\alpha}(V_{\uparrow \text{opt}})(p), \tag{A.32}$$

and therefore that

$$\pi_{\text{opt}}(p) \in \operatorname*{argmax}_{\alpha \in A} \max_{M \in M_{\updownarrow}} VI_{M,\alpha}(V_{\uparrow \text{opt}})(p), \tag{A.33}$$

which implies that $\alpha = \pi_{\text{opt}}(p) \in \rho_{V_{\uparrow \text{opt}}}(p)$. We can then use our subcase assumption that there must be an MDP $M_\alpha \in M_{\updownarrow}$ such that $VI_{M_\alpha,\pi_{\text{opt}}}(V_{\downarrow \text{opt}})(p) < V_{\downarrow \text{opt}}(p)$.

Let M_7 be a π_{opt}-minimizing MDP, as per Theorem 7. Then $V_{M_7,\pi_{\text{opt}}} = V_{\downarrow \pi_{\text{opt}}} = V_{\downarrow \text{opt}}$ by expanding definitions. So $VI_{M_7,\pi_{\text{opt}}} = (V_{\downarrow \pi_{\text{opt}}}) = V_{\downarrow \text{opt}}$. We can now create a new MDP M_8 by copying M_7 at every state except p, where M_8 copies M_α, following the construction used to define M_3 in Subcase 1(a). By construction we then have

$$VI_{M_8,\pi_{\text{opt}}}(V_{\downarrow \text{opt}}) <_{\text{dom}} V_{\downarrow \text{opt}}, \tag{A.34}$$

which by Theorem 6 implies $V_{\downarrow \pi_{\text{opt}}} <_{\text{dom}} V_{\downarrow \text{opt}}$, contradicting our choice of π_{opt} and concluding Subcase 2(b), Case 2, and the proof of Theorem 9. □

Theorem 10. *For any policy π, $IVI_{\downarrow \pi}$ and $IVI_{\uparrow \pi}$ are contraction mappings.*

Proof. We first show that $IVI_{\uparrow \pi}$ is a contraction mapping on $\overline{V}$, the space of value functions. Strictly speaking, $IVI_{\uparrow \pi}$ is a mapping from an interval value function $V_{\updownarrow}$ to a value function V. However, the specific values $V(p)$ only depend on the upper bounds $V_{\uparrow}$ of $V_{\updownarrow}$. Therefore, the mapping $IVI_{\uparrow \pi}$ is isomorphic to a function that maps value functions to value functions and with some abuse of terminology, we can consider $IVI_{\uparrow \pi}$ to be such a mapping. The same is true for $IVI_{\downarrow \pi}$, which depends only on the lower bounds $V_{\downarrow}$.

Let $\hat{u}$ and $\hat{v}$ be interval value functions, fix $p \in Q$, and assume that $IVI_{\uparrow \pi}(\hat{v})(p) \geqslant IVI_{\uparrow \pi}(\hat{u})(p)$. Let M be an MDP $M \in M_{\updownarrow}$ that maximizes the expression $VI_{M,\pi}(v_{\uparrow})(p)$ (Lemma 1 implies that there is such an MDP in the finite set $X_{M_{\updownarrow}}$, guaranteeing the existence of M in spite of the infinite cardinality of $M_{\updownarrow}$).

Then,

$$0 \leqslant IVI_{\uparrow_\pi}(\hat{v})(p) - IVI_{\uparrow_\pi}(\hat{u})(p) \tag{A.35}$$

$$= \max_{M \in M_\updownarrow} VI_{M,\pi}(v_\uparrow)(p) - \max_{M \in M_\updownarrow} VI_{M,\pi}(u_\uparrow)(p) \tag{A.36}$$

$$\leqslant R(p) + \gamma\left(\sum_{q \in Q} F^M_{pq}\big(\pi(p)\big)v_\uparrow(q)\right) - R(p) - \gamma\left(\sum_{q \in Q} F^M_{pq}\big(\pi(p)\big)u_\uparrow(q)\right) \tag{A.37}$$

$$= \gamma\left(\sum_{q \in Q} F^M_{pq}\big(\pi(p)\big)\big[v_\uparrow(q) - u_\uparrow(q)\big]\right) \tag{A.38}$$

$$\leqslant \gamma\left(\sum_{q \in Q} F^M_{pq}\big(\pi(p)\big)\|v_\uparrow - u_\uparrow\|\right) \tag{A.39}$$

$$= \gamma\|v_\uparrow - u_\uparrow\|. \tag{A.40}$$

Line (A.36) expands the definition of $IVI_{\uparrow_\pi}$. Line (A.37) follows by expanding the definition of VI and from the fact that M maximizes $VI_{M,\pi}(v_\uparrow)(p)$ by definition. In line (A.38), we simplify the expression by cancelling the immediate reward terms and factoring out the coefficients F^M_{pq}. In line (A.39), we introduce an inequality by replacing the term $v_\uparrow(q) - u_\uparrow(q)$ with the maximum difference over all states, which by definition is the sup norm. The final step line (A.40) follows from the fact that F is a probability distribution that sums to 1 and $\|v_\uparrow - u_\uparrow\|$ does not depend on q.

Repeating this argument interchanging the roles of $\hat{u}$ and $\hat{v}$ in the case that $IVI_{\uparrow_\pi}(\hat{v})(p) \leqslant IVI_{\uparrow_\pi}(\hat{u})(p)$ implies

$$\big|IVI_{\uparrow_\pi}(\hat{v})(p) - IVI_{\uparrow_\pi}(\hat{u})(p)\big| \leqslant \gamma\|v_\uparrow - u_\uparrow\| \tag{A.41}$$

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

The proof that $IVI_{\downarrow_\pi}$ is a contraction mapping is very similar, replacing $IVI_{\uparrow_\pi}$ with $IVI_{\downarrow_\pi}$ throughout, replacing maximization with minimization in line (A.35), and selecting MDP M to minimize the expression $VI_{M,\pi}(u_\uparrow)(p)$ when $IVI_{\downarrow_\pi}(\hat{v})(p) \geqslant IVI_{\downarrow_\pi}(\hat{u})(p)$. □

Theorem 11. *For any policy π, $V_{\downarrow_\pi}$ is a fixed point of $IVI_{\downarrow_\pi}$ and $V_{\uparrow_\pi}$ of $IVI_{\uparrow_\pi}$, and therefore $V_{\updownarrow_\pi}$ is a fixed point of $IVI_{\updownarrow_\pi}$.*

Proof. We prove the theorem for $IVI_{\downarrow_\pi}$; the proof for $IVI_{\uparrow_\pi}$ is similar. We show

(a) $IVI_{\downarrow_\pi}(V_{\updownarrow_\pi}) \leqslant_{\text{dom}} V_{\downarrow_\pi}$, and

(b) $IVI_{\downarrow_\pi}(V_{\updownarrow_\pi}) \geqslant_{\text{dom}} V_{\downarrow_\pi}$,

from which we conclude that $IVI_{\downarrow_\pi}(V_{\updownarrow_\pi}) = V_{\downarrow_\pi}$. Throughout both cases we take M^* to be a π-minimizing MDP, so that $V_{\downarrow_\pi} = V_{M^*,\pi}$. By Theorem 7 M^* must exist.

We first prove (a). From Theorem 3, we know that $V_{M^*,\pi}$ is a fixed point of $VI_{M^*,\pi}$. Thus, for any state $q \in Q$,

$$V_{\downarrow_\pi}(q) = V_{M^*,\pi}(q) = VI_{M^*,\pi}(V_{M^*,\pi})(q) = VI_{M^*,\pi}(V_{\downarrow_\pi})(q). \tag{A.42}$$

Using this fact and expanding the definition of $IVI_{\downarrow_\pi}$, we have, at every state q,

$$\begin{aligned} IVI_{\downarrow_\pi}(V_{\updownarrow_\pi})(q) &= \min_{M\in M_\updownarrow} VI_{M,\pi}(V_{\downarrow_\pi})(q) \\ &\leqslant VI_{M^*,\pi}(V_{\downarrow_\pi})(q) \\ &= V_{\downarrow_\pi}(q). \end{aligned} \tag{A.43}$$

This implies that $IVI_{\downarrow_\pi}(V_{\updownarrow_\pi}) \leqslant_{\text{dom}} V_{\downarrow_\pi}$ as desired.

To prove (b), suppose for sake of contradiction that for some state p, $IVI_{\downarrow_\pi}(V_{\updownarrow_\pi})(p) < V_{\downarrow_\pi}(p)$. Let $M_1 \in M_\updownarrow$ be an MDP that minimizes [6] the expression $VI_{M_1,\pi}(V_{\downarrow_\pi})(p)$.

Then, substituting M_1 into the definition of $IVI_{\downarrow_\pi}$,

$$IVI_{\downarrow_\pi}(V_{\updownarrow_\pi})(p) = VI_{M_1,\pi}(V_{\downarrow_\pi})(p) < V_{\downarrow_\pi}(p). \tag{A.44}$$

We can then construct an MDP M_2 by copying M^* at every state except p, where M_2 copies M_1 (see the proof of Theorem 9, Case 1(a) for the details of a similar construction). Because M_2 is a copy of M^* at every state but p, Eq. (A.42) must hold with M_2 replacing M^* at every state but p. Because M_2 is a copy of M_1 at state p, Eq. (A.44) with M_2 replacing M_1 must hold at state p. These two facts together imply

$$VI_{M_2,\pi}(V_{\downarrow_\pi}) <_{\text{dom}} V_{\downarrow_\pi}. \tag{A.45}$$

Then by Theorem 6 $V_{M_2,\pi} <_{\text{dom}} V_{\downarrow_\pi}$, contradicting the definition of $V_{\downarrow_\pi}$. □

Theorem 13.

(a) *$IVI_{\uparrow_{\text{opt}}}$ and $IVI_{\downarrow_{\text{pes}}}$ are contraction mappings.*

(b) *For any value function V and associated action set selection function ρ_V and σ_V, $IVI_{\downarrow_{\text{opt},V}}$ and $IVI_{\uparrow_{\text{pes,V}}}$ are contraction mappings.*

Proof. We first prove (a). The proof that $IVI_{\uparrow_{\text{opt}}}$ is a contraction mapping is an extension of the proof of Theorem 10. Let $\hat{u}$ and $\hat{v}$ be interval value functions, fix $p \in Q$, and assume that $IVI_{\uparrow_{\text{opt}}}(\hat{v})(p) \geqslant IVI_{\uparrow_{\text{opt}}}(\hat{u})(p)$. Select $M \in M_\updownarrow$ and $\alpha \in A$ to maximize the expression $VI_{M,\alpha}(v_\uparrow)(p)$ (again, Lemma 1 implies that there is such an MDP in the finite set $X_{M_\updownarrow}$, guaranteeing the existence of M in spite of the infinite cardinality of $M_\updownarrow$).

Then,

$$0 \leqslant IVI_{\uparrow_{\text{opt}}}(\hat{v})(p) - IVI_{\uparrow_{\text{opt}}}(\hat{u})(p) \tag{A.46}$$

$$= \max_{\alpha\in A}\max_{M\in M_\updownarrow} VI_{M,\alpha}(v_\uparrow)(p) - \max_{\alpha\in A}\max_{M\in M_\updownarrow} VI_{M,\alpha}(u_\uparrow)(p) \tag{A.47}$$

$$\leqslant R(p) + \gamma\left(\sum_{q\in Q} F^M_{pq}(\alpha)v_\uparrow(q)\right) - R(p) - \gamma\left(\sum_{q\in Q} F^M_{pq}(\alpha)u_\uparrow(q)\right) \tag{A.48}$$

$$\leqslant \gamma\|v_\uparrow - u_\uparrow\|. \tag{A.49}$$

Line (A.47) expands the definition of $IVI_{\uparrow_{\text{opt}}}$, noting that maximizing using $\leqslant_{\text{opt}}$ selects interval upper bounds based only on the upper bounds of the input intervals. Line (A.48)

[6] Such an MDP exists by Lemma 1, which implies that there must be such an MDP in the finite set $X_{M_\updownarrow} \subseteq M_\updownarrow$.

follows from our choice of M and α to maximize $VI_{M,\alpha}(v_\uparrow)(p)$. Line (A.49) follows from line (A.48) in the same manner that line (A.40) followed from line (A.37) in the proof of Theorem 10, and the desired result for $IVI_{\updownarrow_{\text{opt}}}$ for part (a) of the theorem also follow in the same manner as the remainder of Theorem 10 followed from line (A.40).

To prove that $IVI_{\downarrow_{\text{pes}}}$ is a contraction mapping, we again fix a state p and assume

$$IVI_{\downarrow_{\text{pes}}}(\hat{v})(p) \geqslant IVI_{\downarrow_{\text{pes}}}(\hat{u})(p).$$

We then use $v_\downarrow$ to choose an action α that maximizes $\min_{M\in M_\updownarrow}(VI_{M,\alpha}(v_\downarrow)(p))$ and $u_\downarrow$ to choose an MDP M that minimizes $VI_{M,\alpha}(u_\downarrow)(p)$ (again, Lemma 1 implies that there is such an MDP in the finite set $X_{M_\updownarrow}$, guaranteeing the existence of M). Using α and M as defined above, we have

$$0 \leqslant IVI_{\downarrow_{\text{pes}}}(\hat{v})(p) - IVI_{\downarrow_{\text{pes}}}(\hat{u})(p) \tag{A.50}$$

$$= \max_{\alpha\in A} \min_{M\in M_\updownarrow} VI_{M,\alpha}(v_\downarrow)(p) - \max_{\alpha\in A} \min_{M\in M_\updownarrow} VI_{M,\alpha}(u_\downarrow)(p) \tag{A.51}$$

$$\leqslant \min_{M\in M_\updownarrow} VI_{M,\alpha}(v_\downarrow)(p) - \min_{M\in M_\updownarrow} VI_{M,\alpha}(u_\downarrow)(p) \tag{A.52}$$

$$\leqslant VI_{M,\alpha}(v_\downarrow)(p) - VI_{M,\alpha}(u_\downarrow)(p). \tag{A.53}$$

Line (A.51) expands the definition of $IVI_{\downarrow_{\text{pes}}}$, using the fact that maximizing over $\leqslant_{\text{pes}}$ selects lower bounds based only on the lower bounds of the intervals being maximized over. Line (A.52) substitutes the action α, which introduces the inequality since α was chosen to guarantee

$$\min_{M\in M_\updownarrow} VI_{M,\alpha}(v_\downarrow)(p) = \max_{\alpha\in A} \min_{M\in M_\updownarrow} VI_{M,\alpha}(v_\downarrow)(p), \tag{A.54}$$

and the meaning of maximization guarantees that

$$\min_{M\in M_\updownarrow} VI_{M,\alpha}(u_\downarrow)(p) \leqslant \max_{\alpha\in A} \min_{M\in M_\updownarrow} VI_{M,\alpha}(u_\downarrow)(p). \tag{A.55}$$

Line (A.53) follows similarly because M was chosen to guarantee

$$VI_{M,\alpha}(u_\downarrow)(p) = \min_{M\in M_\updownarrow} VI_{M,\alpha}(u_\downarrow)(p), \tag{A.56}$$

and the meaning of minimization guarantees that

$$VI_{M,\alpha}(v_\downarrow)(p) \geqslant \min_{M\in M_\updownarrow} VI_{M,\alpha}(v_\downarrow)(p). \tag{A.57}$$

The desired result for $IVI_{\updownarrow_{\text{pes}}}$ in part (a) of the theorem then follows directly from line (A.53) in the same manner as the result for $IVI_{\updownarrow_{\text{opt}}}$ followed from line (A.47), concluding the proof of part (a) of the theorem.

For part (b), the proof for $IVI_{\downarrow_{\text{opt},V}}$ follows exactly as the proof for $IVI_{\downarrow_{\text{pes}}}$, except that the set of actions considered in the maximization over actions at each state p is restricted to $\rho_V(p)$. Likewise, proving $IVI_{\uparrow_{\text{pes},V}}$ is the same as proving $IVI_{\uparrow_{\text{opt}}}$ where the set of actions is restricted to $\sigma_V(p)$. □

References

[1] R. Bellman, Dynamic Programming, Princeton University Press, Princeton, NJ, 1957.
[2] D.P. Bertsekas, Dynamic Programming: Deterministic and Stochastic Models, Prentice-Hall, Englewood Cliffs, NJ, 1987.
[3] D.P. Bertsekas, D.A. Castañon, Adaptive aggregation for infinite horizon dynamic programming, IEEE Trans. Autom. Control 34 (6) (1989) 589–598.
[4] D.P. Bertsekas, J.N. Tsitsiklis, Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.
[5] C. Boutilier, T.L. Dean, S. Hanks, Decision theoretic planning: Structural assumptions and computational leverage, J. Artificial Intelligence Res. 11 (1999) 1–94.
[6] C. Boutilier, R. Dearden, Using abstractions for decision theoretic planning with time constraints, in: Proc. AAAI-94, Seattle, WA, 1994, pp. 1016–1022.
[7] C. Boutilier, T. Dean, S. Hanks, Planning under uncertainty: Structural assumptions and computational leverage, in: Proc. 3rd European Workshop on Planning, Assisi, Italy, 1995.
[8] C. Boutilier, R. Dearden, M. Goldszmidt, Exploiting structure in policy construction, in: Proc. IJCAI-95, Montreal, Quebec, 1995, pp. 1104–1111.
[9] T. Dean, R. Givan, Model minimization in Markov decision processes, in: Proc. AAAI-97, Providence, RI, 1997.
[10] T. Dean, R. Givan, S. Leach, Model reduction techniques for computing approximately optimal solutions for Markov decision processes, in: Proc. 13th Conference on Uncertainty in Artificial Intelligence, Providence, RI, 1997, pp. 124–131.
[11] R.A. Howard, Dynamic Programming and Markov Processes, MIT Press, Cambridge, MA, 1960.
[12] M.L. Littman, T.L. Dean, L.P. Kaelbling, On the complexity of solving Markov decision problems, in: Proc. 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95), Montreal, Québec, 1995.
[13] W.S. Lovejoy, Computationally feasible bounds for partially observed Markov decision processes, Oper. Res. 39 (1) (1991) 162–175.
[14] M. Puterman, Markov Decision Processes—Discrete Stochastic Dynamic Programming, Wiley, New York, 1994.
[15] J.K. Satia, R.E. Lave, Markovian decision processes with uncertain transition probabilities, Oper. Res. 21 (1978) 728–740.
[16] L.S. Shapley, Stochastic games, Proc. National Academy of Sciences of the United States of America 39 (1953) 1095–1100.
[17] C.C. White, H.K. Eldeib, Parameter imprecision in finite state, finite action dynamic programs, Oper. Res. 34 (1986) 120–129.
[18] C.C. White, H.K. Eldeib, Markov decision processes with imprecise transition probabilities, Oper. Res. 43 (1994) 739–749.