# A Contracting Dynamical System Perspective toward Interval Markov Decision Processes

Saber Jafarpour and Samuel Coogan

***Abstract*— Interval Markov decision processes are a class of Markov models where the transition probabilities between the states belong to intervals. In this paper, we study the problem of efficient estimation of the optimal policies in Interval Markov Decision Processes (IMDPs) with continuous action-space. Given an IMDP, we show that the pessimistic (resp. the optimistic) value iterations, i.e., the value iterations under the assumption of a competitive adversary (resp. cooperative agent), are monotone dynamical systems and are contracting with respect to the $\ell_\infty$-norm. Inspired by this dynamical system viewpoint, we introduce another IMDP, called the action-space relaxation IMDP. We show that the action-space relaxation IMDP has two key features: (i) its optimal value is an upper bound for the optimal value of the original IMDP, and (ii) its value iterations can be efficiently solved using tools and techniques from convex optimization. We then consider the policy optimization problems at each step of the value iterations as a feedback controller of the value function. Using this system-theoretic perspective, we propose an iteration-distributed implementation of the value iterations for approximating the optimal value of the action-space relaxation IMDP.**

## I. Introduction

*Motivation and Problem Statement:* Markov decision process (MDP) is a powerful and classical framework for modeling the stochastic interactions between a system and its environment [1]. The MDP framework has been successfully used to study various problems in dynamic decision-making [2] and reinforcement learning [3]. A fundamental assumption in the MDP framework is that the parameters of the model are known or are learnable. However, in many real-world applications, the model parameters are typically estimated or inferred using data-driven methods and, thus, they are far from accurate. In the literature, several different approaches have been proposed to analyze MDPs with parameter uncertainties. In [4], [5], robust dynamic programming is proposed to study optimal solutions of Markov decision processes with uncertainty in transition probabilities. In [6], a set-valued fixed-point equation is proposed to study the optimal value of Markov decision processes with uncertain reward functions. In [7], computationally efficient algorithms are developed to infer the unknown parameters in MDPs.

Interval Markov decision processes (IMDPs) are a class of Markov models with interval-bounded transition probabilities and reward functions [8]. IMDPs can be considered as a family of MDPS and they appear naturally in the setting where systems are modeled using MDPs with uncertain parameters or with parameters obtained from data-driven sampling approaches. An alternative interpretation for the IMDP framework comes from a game-theoretic perspective. In this case, an IMDP models how an MDP interacts with the environment in the presence of an agent who resolves uncertain transition probabilities [9]. In the literature, IMDPs have been used to analyze a various tasks including checking temporal logic specifications [10] and motion planning in robotics [11]. Much of the early works on the IMDPs focus on models with finite number of actions [8], [4], [5]. Recently, IMDPs with continuous-action spaces have gained attention due to their role in finite-state abstraction of stochastic dynamical systems [12], [13] and in reachability analysis of stochastic systems [14]. Value iterations for IMDPs with continuous action-spaces are studied in [9] and in [15].

One of the main challenges in studying IMDPs with continuous action-space arises in computing their optimal policies. It turns out that most of the existing iterative algorithms for estimating optimal policy of MDPs and IMDPs (including value iteration, policy-iteration, and their interval-valued counterparts) require solving an optimization problem in the action variables at each iteration step. Unlike finite-state finite-action IMDPs where the optimization over the action-space can be implemented efficiently, for IMDPs with continuous action-spaces, optimization over action variables can lead to two important challenges. First, in the absence of any structure for the optimization problem (*e.g.* convexity/concavity of the cost function), it is generally necessary to resort to heuristic algorithms to approximate the solutions of these optimization problems. These heuristic methods can significantly degrade the quality of the estimated optimal policies and can ruin any guarantee on the optimality of the solutions. Secondly, in large-scale IMDPs, solving these optimization problems at each iteration step is computationally complicated, potentially leading to intractability of finding the optimal values. Most of the existing literature on IMDPs focuses on discretizing the action-space and then using known results about discrete-action IMDPs. However, this approximation is sub-optimal and scales poorly with the dimension of the action-space [9]. The only exception is [13] which provides a computationally efficient reformulation of interval value iterations.

*Contributions:* In this paper, we study IMDPs with continuous action-spaces from a dynamical system perspective.

*This work is partially supported by the National Science Foundation under awards #1749357, #1931980, and #1924978, the Air Force Office of Scientific Research under Grant FA9550-23-1-0303, and NASA under the University Leadership Initiative (ULI) Award #80NSSC20M0163 but solely reflects the opinions and conclusions of its authors.

Saber Jafarpour and Samuel Coogan are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, USA, {saber,sam.coogan}@gatech.edu

In particular, we use monotone system theory and contraction theory to study convergence of their value iterations for both pessimistic and optimistic policies, that is, policies under the assumption of a adversarial agent and a cooperative agent, respectively. By considering the value iterations in IMDPs as dynamical systems, we study contractivity and monotonicity of the pessimistic and optimistic value iterations with respect to the $\ell_\infty$-norm and the standard partial order. Next, given an IMDP, we introduce another IMDP, called the action-space relaxation IMDP, obtained by bounding its probability transition and rewards using suitable convex/concave functions. As our first main result, we use a dynamical systems perspective to show that the optimal value of action-space relaxation IMDP provides bound on the optimal value of the original MDP. As our second main result, given an action-space relaxation IMDP, we propose to reduce the computational burden of the interval value iterations by implementing the policy optimization problem in an iteration-distributed fashion. We consider the value iteration and the policy optimization problem as an interconnected dynamical system and leverage the contractivity of the value iterations to provide guarantees for convergence of the interconnected system to the optimal value of the action-space relaxation IMDP.

## II. Notations and Mathematical Preliminary

For every $p \in [1, \infty]$, we denote the $\ell_p$-norm on $\mathbb{R}^n$ by $\|\cdot\|_p$. Given two sets $X$ and $Y$, the set of all the maps from $X$ to $Y$ is denoted by $Y^X$. For any compact set $X \subseteq \mathbb{R}^n$, we define $\|X\|_\infty = \{\|x - y\|_\infty \mid x, y \in X\}$. Let $S$ be a finite set with $n$ elements and let $v \in \mathbb{R}^S$. We define $\{1_v, \dots, n_v\}$ as an ordered permutation of elements of the set $S$ such that $v(1_v) \geq v(2_v) \geq \dots \geq v(n_v)$. The set of all compact interval subsets of $[a, b]$ is denoted by $\mathrm{Interval}_{[a,b]}$ and the set of all compact interval subsets of $\mathbb{R}$ is denoted by $\mathrm{Interval}_{\mathbb{R}}$, i.e., we have

$$\mathrm{Interval}_{[a,b]} = \{[x, y] \mid a \leq x \leq y \leq b\}$$
$$\mathrm{Interval}_{\mathbb{R}} = \{[x, y] \mid x \leq y\}.$$

Given an operator $f : \mathbb{R}^n \to \mathbb{R}^n$, we say that $f$ is *monotone* if, for every $x \leq y$, we have $f(x) \leq f(y)$. Given an operator $f : \mathbb{R}^n \to \mathbb{R}^n$, we say that $f$ is *monotone* if, for every $x \leq y$, we have $f(x) \leq f(y)$. Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, we say that $f$ is *contracting* with rate $\lambda \in (0, 1)$ with respect to the norm $\|\cdot\|$, if

$$\|f(x) - f(y)\| \leq \lambda \|x - y\|, \quad \text{for every } x, y \in \mathbb{R}^n.$$

Given a compact set $\mathcal{X} \subseteq \mathbb{R}^n$, the orthogonal projection into $\mathcal{X}$ is denoted by $\mathrm{Proj}_{\mathcal{X}}$, i.e., $\mathrm{Proj}_{\mathcal{X}}(y) = \mathrm{argmin}_{x \in \mathcal{X}} \|x - y\|_2$, for every $y \in \mathbb{R}^n$. We also recall the setting of a discounted infinite-horizon Markov Decision Process (MDP) with continuous action-space. An MDP with continuous action-space is a tuple $\mathcal{M} = (S, A, P, R, \gamma)$ where

(i) $S$ is a finite set of states.
(ii) $A \subseteq \mathbb{R}^m$ is a compact action space.
(iii) $P : S \times S \times A \to [0, 1]$ is the transition probability function, i.e., for every $s \in S$ and every $a \in A$, $P(s', s, a)$ is the probability of arriving at state $s'$ by taking action $a$ in the state $s$. We assume $0 \leq P(s', s, a) \leq 1$ for every $s, s' \in S$ and every $a \in A$ and we have $\sum_{s' \in S} P(s', s, a) = 1$.
(iv) $R : S \times A \to \mathbb{R}$ is the reward function where $R(s, a)$ is the cost of taking action $a$ at state $s$.
(v) $\gamma \in (0, 1)$ is a discount factor.

A *policy* for the MDP $\mathcal{M}$ is a vector $\pi \in A^S$ which assigns an action $a$ to each state $s$[1]. For every policy $\pi \in A^S$, we define the value function $V_\pi^{\mathcal{M}} : S \to \mathbb{R}$ as

$$V_\pi^{\mathcal{M}}(s) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) \Big| s_0 = s\right) \tag{1}$$

where $\{s_t\}_{t=0}^{\infty}$ is a time sequence of states starting from $s_0 = s$ and following the policy $\pi$. The goal is to find a policy $\pi^* \in A^S$ which maximizes the value function $V_\pi^{\mathcal{M}}$, i.e., a policy $\pi^* \in A^S$ such that

$$\pi^* = \mathrm{argmax}_{\pi \in A^S} V_\pi^{\mathcal{M}}. \tag{2}$$

In general, it can be shown that the optimization problem (2) has a unique optimal value $V^*$, which is obtained at a policy $\pi^* \in A^S$ [1, Theorem 6.1.1], i.e., $V_{\pi^*}^{\mathcal{M}} = V^*$. It can be shown that the optimal value $V^*$ satisfies

$$V^*(s) = R(s, \pi^*(s)) + \gamma \sum_{s' \in S} P(s', s, \pi^*(s)) V^*(s').$$

The *Bellman-policy operator* $\mathsf{F} : \mathbb{R}^S \times A^S \to \mathbb{R}^S$ is

$$\mathsf{F}(v, \pi)(s) := R(s, \pi(s)) + \gamma \sum_{s' \in S} P(s', s, \pi(s)) v(s') \tag{3}$$

and the *Bellman operator* $\mathsf{G} : \mathbb{R}^S \to \mathbb{R}^S$ is defined by

$$\mathsf{G}(v)(s) := \max_{a \in A} \left\{ R(s, a) + \gamma \sum_{s' \in S} P(s', s, a) v(s') \right\}. \tag{4}$$

Equivalently, using the vector notation, we have $\mathsf{G}(v) = \max_{\pi \in A^S} \mathsf{F}(v, \pi)$, for every $v \in \mathbb{R}^S$. It is known that the Bellman operator $\mathsf{G}$ is contracting with respect to the $\ell_\infty$-norm, monotone with respect to the standard partial ordering, and the optimal value $V^*$ is the fixed point of the Bellman operator, i.e., $V^* = \mathsf{G}(V^*)$ [1, Theorem 6.2.3].

## III. Interval Markov Decision Process

In this section, we introduce *Interval Markov Decision Processes (IMDPs)* as a class of Markov models where the cost functions and probability transitions are unknown and belong to suitable intervals. An IMDP is a tuple $\mathcal{IM} = (S, A, [P], [R], \gamma)$ where

(i) $S$ is a finite set of states.
(ii) $A \subseteq \mathbb{R}^m$ is a compact action space.
(iii) $[P] \in S \times S \times A \to \mathrm{Interval}_{[0,1]}$ denotes the transition probability intervals, i.e., for every $s \in S$ and every $a \in A$, $[P](s', s, a) = [\underline{P}(s', s, a), \overline{P}(s', s, a)]$ is the probability interval of arriving to the state $s'$ by taking action

[1]A policy defined this way is usually referred to as a Markovian deterministic stationary policy in the literature [1].

$a$ in the state $s$. For the sake of consistency, we assume that $\sum_{s' \in S} \underline{P}(s', s, a) \leq 1 \leq \sum_{s' \in S} \overline{P}(s', s, a)$.

(iv) $[R] : S \times A \to \text{Interval}_{\mathbb{R}}$ denotes the reward function where $[R](s, a) = [\underline{R}(s, a), \overline{R}(s, a)]$ is the reward interval of taking action $a$ at state $s$.

(v) $\gamma \in (0, 1)$ is a discount factor.

An MDP $\mathcal{M} = (S, A, P, R, \gamma)$ belongs to the IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$, and we write $\mathcal{M} \in \mathcal{IM}$, if

$$\underline{P}(s', s, a) \leq P(s', s, a) \leq \overline{P}(s', s, a),$$
$$\underline{R}(s, a) \leq R(s, a) \leq \overline{R}(s, a),$$

for every $s, s' \in S$ and every $a \in A$. A policy for $\mathcal{IM}$ is a vector $\pi \in A^S$ that assigns an action $a$ to each state $s$.

*Remark 3.1: (Comparison with the literature)* Our definition of IMDPs generalizes the classical definitions in [8], [5] which assume finite action-spaces. This generalization is motivated by, *e.g.,* applications in robotics [11] and abstraction of stochastic dynamical systems [14], [9]. Moreover, two different interpretations for IMDPs have been proposed in the literature. The first interpretation considers an IMDP as an MDP with uncertain parameters [8], [5], whereas the second interpretation considers an IMDP as an MDP interacting with an agent who resolves uncertain probabilities [6], [9].

Given an IMDP $\mathcal{IM}$ and a policy $\pi \in A^S$, we study the possible ranges of the value function (1) for every $\mathcal{M} \in \mathcal{IM}$. First, for every $(s, a) \in S \times A$, we define $\Delta^{\mathcal{IM}}_{s,a}$ as the set of all $p : S \to [0, 1]$ such that,

$$\underline{P}(s', s, a) \leq p(s') \leq \overline{P}(s', s, a), \text{ for all } s' \in S$$
$$\sum\nolimits_{r \in S} p(r) = 1. \tag{5}$$

Using the set $\Delta^{\mathcal{IM}}_{s,a}$, we define the *interval Bellman-policy operator* for $\mathcal{IM}$ as the map $\begin{bmatrix} \underline{\mathsf{F}} \\ \overline{\mathsf{F}} \end{bmatrix} : \mathbb{R}^S \times A^S \to \mathbb{R}^S \times \mathbb{R}^S$:

$$\underline{\mathsf{F}}(v, \pi)(s) = \underline{R}(s, \pi(s)) + \gamma \min_{p \in \Delta^{\mathcal{IM}}_{s,\pi(s)}} \sum_{s' \in S} p(s')v(s'),$$
$$\overline{\mathsf{F}}(v, \pi)(s) = \overline{R}(s, \pi(s)) + \gamma \max_{p \in \Delta^{\mathcal{IM}}_{s,\pi(s)}} \sum_{s' \in S} p(s')v(s'), \tag{6}$$

and the *interval Bellman operator* for $\mathcal{IM}$ as the map $\begin{bmatrix} \underline{\mathsf{G}} \\ \overline{\mathsf{G}} \end{bmatrix} : \mathbb{R}^S \to \mathbb{R}^S \times \mathbb{R}^S$:

$$\underline{\mathsf{G}}(v)(s) = \max_{a \in A} \left\{ \underline{R}(s, a) + \gamma \min_{p \in \Delta^{\mathcal{IM}}_{s,a}} \sum_{s' \in S} p(s')v(s') \right\},$$
$$\overline{\mathsf{G}}(v)(s) = \max_{a \in A} \left\{ \overline{R}(s, a) + \gamma \max_{p \in \Delta^{\mathcal{IM}}_{s,a}} \sum_{s' \in S} p(s')v(s') \right\}. \tag{7}$$

Equivalently, using the vector notation, we have $\underline{\mathsf{G}}(v) = \max_{\pi \in A^S} \underline{\mathsf{F}}(v, \pi)$ and $\overline{\mathsf{G}}(v) = \max_{\pi \in A^S} \overline{\mathsf{F}}(v, \pi)$. for every $v \in \mathbb{R}^S$. In the next theorem, we show that the interval Bellman (resp. Bellman-policy) operator is monotone and contracting with respect to the $\ell_\infty$-norm and can be used to provide upper and lower bounds on the Bellman (resp. Bellman-policy) operator of every MDP that belongs to $\mathcal{IM}$. The proof of this theorem is provided in [16, Appendix A].

*Theorem 3.2 (Interval Bellman operator):* Consider an IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$ with the interval Bellman-policy operator and the interval Bellman operator in (6) and (7), respectively. Let $\mathcal{M}$ be an MDP such that $\mathcal{M} \in \mathcal{IM}$ with the Bellman-policy operator and the Bellman operator operator in (3) and (4), respectively. Then,

(i) for every $\pi \in A^S$, the operators $v \mapsto \underline{\mathsf{F}}(v, \pi)$ and $v \mapsto \overline{\mathsf{F}}(v, \pi)$ are monotone and contracting with respect to the $\ell_\infty$-norm with rate $\gamma$ and

$$\underline{\mathsf{F}}(v, \pi) \leq \mathsf{F}(v, \pi) \leq \overline{\mathsf{F}}(v, \pi), \quad \text{for all } v \in \mathbb{R}^S. \tag{8}$$

(ii) the operators $\underline{\mathsf{G}}$ and $\overline{\mathsf{G}}$ are monotone and contracting with respect to the $\ell_\infty$-norm with rate $\gamma$ and

$$\underline{\mathsf{G}}(v) \leq \mathsf{G}(v) \leq \overline{\mathsf{G}}(v), \quad \text{for all } v \in \mathbb{R}^S. \tag{9}$$

Computing the interval Bellman operator using the equation (6) requires solving two linear programs in $p$ and can be computationally intractable for large-scale IMDPs. We first introduce two useful notations. Consider $\mathcal{IM} = (S, A, [P], [R], \gamma)$ with $(s, a) \in S \times A$ and $v \in \mathbb{R}^S$. Then we define $\underline{\iota}(v, s, a)$ as the largest integer $j \in \{1, \ldots, n\}$ satisfying

$$\underline{P}(j_v, s, a) \leq 1 - \sum_{i=1}^{j-1} \underline{P}(i_v, s, a) - \sum_{i=j+1}^{n} \overline{P}(i_v, s, a) \leq \overline{P}(j_v, s, a),$$

and $\overline{\iota}(v, s, a)$ as the largest integer $k \in \{1, \ldots, n\}$ satisfying

$$\underline{P}(k_v, s, a) \leq 1 - \sum_{i=1}^{k-1} \overline{P}(i_v, s, a) - \sum_{i=k+1}^{n} \underline{P}(i_v, s, a) \leq \overline{P}(k_v, s, a).$$

Note that existence of $\underline{\iota}(v, s, a)$ and $\overline{\iota}(v, s, a)$ follows from the inequality $\sum_{s' \in S} \underline{P}(s', s, a) \leq 1 \leq \sum_{s' \in S} \overline{P}(s', s, a)$. We also define the operators $\Omega^{\mathcal{IM}} : \mathbb{R}^S \times S \times A \to \mathbb{R}$ and $\Lambda^{\mathcal{IM}} : \mathbb{R}^S \times S \times A \to \mathbb{R}$ as follows:

$$\Omega^{\mathcal{IM}}(v, s, a) = \sum_{i=1}^{j} (v(i_v) - v(j_v)) \underline{P}(i_v, s, a) + \sum_{i=j}^{n} (v(i_v) - v(j_v)) \overline{P}(i_v, s, a) + v(j_v),$$
$$\Lambda^{\mathcal{IM}}(v, s, a) = \sum_{i=1}^{k} (v(i_v) - v(k_v)) \overline{P}(i_v, s, a) + \sum_{i=k}^{n} (v(i_v) - v(k_v)) \underline{P}(i_v, s, a) + v(k_v), \tag{10}$$

where $j = \underline{\iota}(v, s, a)$ and $k = \overline{\iota}(v, s, a)$. The next proposition provides a closed-from expression for the interval Bellman-policy using the lower and upper probability transition bounds. The proof is provided in [16, Appendix B].

*Proposition 3.3 (Bellman-policy operator):* Consider the IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$ with the interval Bellman-policy operator $\begin{bmatrix} \underline{\mathsf{F}} \\ \overline{\mathsf{F}} \end{bmatrix}$ defined in (6). Then

$$\underline{\mathsf{F}}(v, \pi)(s) = \underline{R}(s, \pi(s)) + \gamma \Omega^{\mathcal{IM}}(v, s, \pi(s)),$$
$$\overline{\mathsf{F}}(v, \pi)(s) = \overline{R}(s, \pi(s)) + \gamma \Lambda^{\mathcal{IM}}(v, s, \pi(s)),$$

where $\Omega^{\mathcal{IM}}$ and $\Lambda^{\mathcal{IM}}$ are defined in (10).

It is known that the notion of *optimal policy*, as defined in (2) for MDPs, is not well-defined for IMDPs [8]. This is due to the fact that the value functions of IMDPs are interval-valued and the set of intervals do not have a standard partial order. However, given an IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$, one can define two policies, namely the *pessimistic optimal policy* and the *optimistic optimal policy*, which provide certain type of optimally for the value function. The pessimistic optimal policy $\pi_{\mathrm{p}}^* \in A^S$ is the unique policy defined by

$$\pi_{\mathrm{p}}^* = \mathrm{argmax}_{\pi \in A^S} \left( \min_{\mathcal{M} \in \mathcal{IM}} V_\pi^{\mathcal{M}} \right),$$

and the pessimistic value function is given by $V_{\mathrm{p}}^* = \min_{\mathcal{M} \in \mathcal{IM}} V_{\pi_{\mathrm{p}}^*}^{\mathcal{M}}$. The optimistic optimal policy $\pi_{\mathrm{o}}^* \in A^S$ is the unique policy defined by

$$\pi_{\mathrm{o}}^* = \mathrm{argmax}_{\pi \in A^S} \left( \max_{\mathcal{M} \in \mathcal{IM}} V_\pi^{\mathcal{M}} \right),$$

and the optimistic value function is given by $V_{\mathrm{o}}^* = \max_{\mathcal{M} \in \mathcal{IM}} V_{\pi_{\mathrm{o}}^*}^{\mathcal{M}}$. From a game-theoretic perspective, the pessimistic optimal policy can be considered as the optimal policy of the IMDP $\mathcal{IM}$ in presence of a competitive adversary who resolves uncertain probabilities, and the optimistic optimal policy can be considered as the optimal policy of the IMDP $\mathcal{IM}$ the presence of a cooperative agent who resolves uncertain probabilities. Given an IMDP $\mathcal{IM}$, we define the *pessimistic value iteration* by

$$v^{k+1} = \underline{\mathsf{G}}(v^k) = \max_{\pi \in A^S} \underline{\mathsf{F}}(v^k, \pi), \tag{11}$$

and we define the *optimistic value iteration* by

$$v^{k+1} = \overline{\mathsf{G}}(v^k) = \max_{\pi \in A^S} \overline{\mathsf{F}}(v^k, \pi). \tag{12}$$

where $\begin{bmatrix} \underline{\mathsf{G}} \\ \overline{\mathsf{G}} \end{bmatrix}$ and $\begin{bmatrix} \underline{\mathsf{F}} \\ \overline{\mathsf{F}} \end{bmatrix}$ are the interval Bellman operator and the interval Bellman-policy operator of $\mathcal{IM}$, respectively. The next theorem establishes that the pessimistic and optimistic value iterations (11) and (12) can be used to compute the pessimistic and optimistic optimal policies of IMDPs. We refer to [16, Appendix C] for the proof.

*Theorem 3.4 (Value iterations as dynamical systems):* Consider the IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$ with the pessimistic and optimistic policies $\pi_{\mathrm{p}}^*, \pi_{\mathrm{o}}^* \in A^S$ with the interval Bellman-policy operator (6) and the interval Bellman operator (7). Then, the following statements hold:

(i) the pessimistic value iteration (11) is a monotone contracting dynamical system with the unique globally exponentially stable equilibrium point $V_{\mathrm{p}}^*$ and the pessimistic optimal policy $\pi_{\mathrm{p}}^*$ is obtained by $\pi_{\mathrm{p}}^* = \mathrm{argmax}_{\pi \in A^S} \underline{\mathsf{F}}(V_{\mathrm{p}}^*, \pi)$.

(ii) the optimistic value iteration (12) is a monotone contracting dynamical system with the unique globally exponentially stable equilibrium point $V_{\mathrm{o}}^*$ and the optimistic optimal policy $\pi_{\mathrm{o}}^*$ is obtained by $\pi_{\mathrm{o}}^* = \mathrm{argmax}_{\pi \in A^S} \overline{\mathsf{F}}(V_{\mathrm{o}}^*, \pi)$.

*Remark 3.5 (A dynamical system perspective):* The fact that pessimistic (resp. optimistic) value iterations is contracting and the pessimistic (resp. optimistic) optimal policy is its fixed points is known in the literature [8, Theorems 10,11,12]. However, Theorem 3.4 provides a discrete-time dynamical system perspective to the pessimistic (resp. optimistic) value iterations (11) (resp. equation (12)) and highlights their less-studied property of monotonicity.

## IV. Efficient estimation of optimal policies

Theorem 3.4 provides iterative algorithms for computing the optimal policies in IMDPs. It turns out that implementing the pessimistic value iterations (11) (resp. optimistic value iterations (12)) requires solving the following $|S|$ nonlinear optimization problems at each iteration step:

$$\underline{\pi}^k = \mathrm{argmax}_{\pi \in A^S} \underline{\mathsf{F}}(v^k, \pi) \tag{13}$$

(resp. $\overline{\pi}^k = \mathrm{argmax}_{\pi \in A^S} \overline{\mathsf{F}}(v^k, \pi)$). This can cause two main challenges for computing the optimal policies:

(i) in the absence of any structure for the optimization problems (13), one needs to resort to heuristic algorithms to approximate the optimal solutions of (13). These heuristic algorithms can introduce sizable error in estimating the optimization problem and can significantly degrade the performance of the value iterations.

(ii) even when the optimization problems (13) is convex, it is still necessary to solve $|S|$ optimization problems with $m$ variables at each iterations of the value iterations. Thus, it is computationally challenging to implement the interval value iterations for large-scale IMDPs.

In order to address the above mentioned challenges, we study IMDPs through the lens of dynamical systems. In the rest of this section, we focus on the pessimistic value iterations and pessimistic optimal policies. A parallel framework can be developed for optimistic value iterations and optimistic optimal policies but we omit it for the sake of brevity.

### A. Action-space relaxation IMDP

In this subsection, we introduce a relaxation of a given IMDP in its action variables by providing suitable bounds on its reward functions and its probability transition functions.

*Definition 4.1 (Action-space pessimistic relaxation):* Consider an IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$. An *action-space pessimistic relaxation* of $\mathcal{IM}$ is an IMDP $\mathcal{IM}^{\mathrm{cv}} = (S, A^{\mathrm{cv}}, [P^{\mathrm{cv}}], [R^{\mathrm{cv}}], \gamma)$ such that

(i) $A \subseteq A^{\mathrm{cv}} \subseteq \mathbb{R}^m$ and $A^{\mathrm{cv}}$ is convex and compact in $\mathbb{R}^m$,

(ii) for every $s', s \in S$ and every $a \in A$.

$$\underline{P}(s', s, a) \leq \underline{P}^{\mathrm{cv}}(s', s, a), \quad \overline{P}^{\mathrm{cv}}(s', s, a) \leq \overline{P}(s', s, a),$$
$$\underline{R}(s, a) \leq \underline{R}^{\mathrm{cv}}(s, a),$$

(iii) for every $s \in S$, $a \mapsto \underline{R}^{\mathrm{cv}}(s, a)$ is concave on $A^{\mathrm{cv}}$,

(iv) for every $s', s \in S$, $a \mapsto \overline{P}^{\mathrm{cv}}(s', s, a)$ is convex and $a \mapsto \underline{P}^{\mathrm{cv}}(s', s, a)$ is concave.

Given an action-space pessimistic relaxation $\mathcal{IM}^{\mathrm{cv}}$ for $\mathcal{IM}$, one can define its associated interval Bellman-policy operator $\begin{bmatrix} \underline{\mathsf{F}}^{\mathrm{cv}} \\ \overline{\mathsf{F}}^{\mathrm{cv}} \end{bmatrix} : \mathbb{R}^S \times (A^{\mathrm{cv}})^S \to \mathbb{R}^S \times \mathbb{R}^S$ and the interval Bellman operator $\begin{bmatrix} \underline{\mathsf{G}}^{\mathrm{cv}} \\ \overline{\mathsf{G}}^{\mathrm{cv}} \end{bmatrix} : \mathbb{R}^S \times (A^{\mathrm{cv}})^S \to \mathbb{R}^S \times \mathbb{R}^S$ as in equations (6) and (7), respectively. Then, the pessimist value iterations for $\mathcal{IM}^{\mathrm{cv}}$ is given by

$$\begin{aligned} v^{k+1} &= \underline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi^k), \\ \pi^k &= \operatorname{argmax}_{\pi \in (A^{\mathrm{cv}})^S} \underline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi), \end{aligned} \tag{14}$$

and the pessimistic optimal value and the pessimistic optimal policy of $\mathcal{IM}^{\mathrm{cv}}$ is denoted by $V_{\mathrm{p}}^{\mathrm{cv},*}$ and $\pi_{\mathrm{p}}^{\mathrm{cv},*}$, respectively. Given an IMDP $\mathcal{IM}$, the next theorem shows that the Bellman operator of the action-space pessimistic relaxation $\mathcal{IM}^{\mathrm{cv}}$ is an upper bound for the Bellman operator of $\mathcal{IM}$. Using the classical comparison theorem for the pessimistic value iterations (14), it can be shown that the pessimistic optimal value of $\mathcal{IM}^{\mathrm{cv}}$ is an upper bound for the pessimistic optimal value of $\mathcal{IM}$.

*Theorem 4.2 (Bellman operator of pessimistic relaxation):* Consider the IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$ with an associated action-space pessimistic relaxation IMDP $\mathcal{IM}^{\mathrm{cv}} = (S, A^{\mathrm{cv}}, [P^{\mathrm{cv}}], [R^{\mathrm{cv}}], \gamma)$. Then,

(i) for every $v \in \mathbb{R}^S$ and $\pi \in A^S$, $\underline{\mathsf{F}}(v, \pi) \leq \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)$,
(ii) for every $s \in S$ and every $v \in \mathbb{R}^S$, $\pi \to \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)(s)$, is a concave function on $A^{\mathrm{cv}}$.
(iii) for every $v \in \mathbb{R}^S$, $\underline{\mathsf{G}}(v) \leq \underline{\mathsf{G}}^{\mathrm{cv}}(v)$.
(iv) we have $V_{\mathrm{p}}^* \leq V_{\mathrm{p}}^{\mathrm{cv},*}$.

*Proof:* Regarding part (i), recall the definition of $\Omega^{\mathcal{IM}^{\mathrm{cv}}}(v, s, a)$ in equation (10) for the IMDP $\mathcal{IM}^{\mathrm{cv}}$. Then,

$$\begin{aligned} \Omega^{\mathcal{IM}^{\mathrm{cv}}}(v, s, \pi(s)) &= \sum_{i=1}^{j} (v(i_v) - v(j_v)) \underline{P}^{\mathrm{cv}}(i_v, s, a) \\ &\quad + \sum_{i=j}^{n} (v(i_v) - v(j_v)) \overline{P}^{\mathrm{cv}}(i_v, s, a) + v(j_v) \\ &\geq \sum_{i=1}^{j} (v(i_v) - v(j_v)) \underline{P}(i_v, s, a) \\ &\quad + \sum_{i=j}^{n} (v(i_v) - v(j_v)) \overline{P}(i_v, s, a) + v(j_v), \end{aligned}$$

where $j = \underline{\iota}^{\mathrm{cv}}(v, s, a)$ is as defined in (10). Note that the first inequality above holds because, for every $i \in \{1, \ldots, j\}$, we have $v(i_v) - v(j_v) \geq 0$ and $\underline{P}^{\mathrm{cv}}(i_v, s, a) \geq \underline{P}(i_v, s, a)$ and, for every $i \in \{j, \ldots, n\}$, we have $v(i_v) - v(j_v) \leq 0$ and $\overline{P}^{\mathrm{cv}}(i_v, s, a) \leq \overline{P}(i_v, s, a)$. Given $\pi \in A^S$, we define $p^* : S \to \mathbb{R}$ by

$$p^*(i_v) = \begin{cases} \underline{P}(i_v, s, \pi(s)) & i \in \{1, \ldots, j-1\} \\ \xi & i = j \\ \overline{P}(i_v, s, \pi(s)) & i \in \{j+1, \ldots, n\}. \end{cases}$$

where $\xi = 1 - \sum_{i=1}^{j-1} \underline{P}(i_v, s, \pi(s)) - \sum_{i=j+1}^{n} \overline{P}(i_v, s, \pi(s))$. It is easy to check $p^* \in \Delta^{\mathcal{IM}}_{s,\pi(s)}$ and

$$\Omega^{\mathcal{IM}^{\mathrm{cv}}}(v, s, \pi(s)) \geq \sum_{s' \in S} p^*(s') v(s').$$

As a result, using Proposition 3.3, we get

$$\begin{aligned} \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)(s) &= \underline{R}^{\mathrm{cv}}(s, \pi(s)) + \gamma \Omega^{\mathcal{IM}^{\mathrm{cv}}}(v, s, \pi(s)) \\ &\geq \underline{R}(s, \pi(s)) + \gamma \min_{\pi \in \Delta^{\mathcal{IM}}_{s,\pi(s)}} \sum_{s' \in S} p(s') v(s') = \underline{\mathsf{F}}(v, \pi)(s). \end{aligned}$$

where the last equality holds by the definition of $\underline{\mathsf{F}}$. Regarding part (ii), first note that, for every $i \in \{1, \ldots, j\}$, we have $v(i_v) - v(j_v) \geq 0$ and $a \mapsto \underline{P}^{\mathrm{cv}}(i_v, s, a)$ is concave and, for every $i \in \{j, \ldots, n\}$, we have $v(i_v) - v(j_v) \leq 0$ and $a \mapsto \overline{P}^{\mathrm{cv}}(i_v, s, a)$ is convex. This implies that $\pi \mapsto \Omega^{\mathcal{IM}^{\mathrm{cv}}}(v, s, \pi(s))$ is an concave function. Moreover,

$$\underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)(s) = \underline{R}^{\mathrm{cv}}(s, \pi(s)) + \gamma \Omega^{\mathcal{IM}^{\mathrm{cv}}}(v, s, \pi(s))$$

Since $a \mapsto \underline{R}^{\mathrm{cv}}(s, a)$ is concave, we can deduce that $\pi \mapsto \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)(s)$ is a concave function. Regarding part (iii), the fact that $\underline{\mathsf{G}}(v) \leq \underline{\mathsf{G}}^{\mathrm{cv}}(v)$ follows from definition of $\underline{\mathsf{G}}$ in (7). Regarding part (iv), by Theorem 3.4(i), the discrete-time dynamical systems (11) and (14) are monotone and contracting with respect to $\ell_\infty$-norm. Note that $\underline{\mathsf{G}}(v) \leq \underline{\mathsf{G}}^{\mathrm{cv}}(v)$, for every $v \in \mathbb{R}^S$. Therefore, we can use the comparison theorem [17, Theorem 3.8.1], to get $V_{\mathrm{p}}^* \leq V_{\mathrm{p}}^{\mathrm{cv},*}$. ■

*Remark 4.3:* The following remarks are in order.

(i) *(Computational efficiency)*: using the action-space pessimistic relaxation $\mathcal{IM}^{\mathrm{cv}}$, Theorem 4.2 develops the iteration scheme (14) for over-approximating the pessimistic optimal value function of the original IMDP $\mathcal{IM}$. Since the interval Bellman-policy operator is concave in $\pi$, standard convex optimization algorithms (see [18]) can be employed to solve the optimization problem at each iteration of (14).
(ii) *(Novelty)*: to the best of our knowledge, [9] is the first paper that proposes to use the concave/convex bounds on the parameters of the IMDPs to approximate its optimal policies. Compared to [9], Definition 4.1 and Theorem 4.2 develop a rigorous framework to bound the parameters of the IMDP and provide guarantees for over-approximation of their optimal values. Moreover, our framework is capable of dealing with IMDPs with action-dependent reward functions.

### B. Iteration-distributed optimization

In practice, estimating the optimal policies using the pessimistic value iterations (14) requires solving $|S|$ concave optimization problems with $m$ variables at each iteration step, which can become computationally intractable for IMDPs with large state-space. In this subsection, we consider the pessimistic value iterations (14) as the interconnection of a dynamical system described by the value iterations:

$$v^{k+1} = \underline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi^k), \tag{15}$$

and an optimization-based feedback controller described by:

$$\pi^k = \operatorname{argmax}_{\pi\in(A^{\mathrm{cv}})^S} \underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi). \tag{16}$$

Using this system-theoretic perspective toward interval value iterations (14), we propose to implement the optimization-based feedback controller in a distributed fashion. We first need to introduce the following assumption on the IMDPs.

*Assumption 4.4:* For the IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$,

(i) *(Bounded rewards):* there exists $\underline{m} \in \mathbb{R}_{\geq 0}$ such that $\sup_{s\in S, a\in A} \underline{R}(s,a) \leq \underline{m}$,

(ii) *(Regularity in action variables):* the map $a \mapsto \underline{R}(s,a)$ is twice continuously differentiable and $c$-strongly concave and $L$-smooth, uniformly in $s \in S$, and the maps

$$a \mapsto \underline{P}(s',s,a), \qquad a \mapsto \overline{P}(s',s,a),$$

are twice continuously differentiable and $L$-smooth, for every $s', s \in S$.

Given an IMDP $\mathcal{IM}$ with an action-space pessimistic relaxation $\mathcal{IM}^{\mathrm{cv}}$, we replace the feedback controller described by the optimization problem (16) with $\ell \in \mathbb{Z}_{\geq 0}$ iteration of the projected gradient descent operator $\underline{\mathsf{T}}^{\mathrm{cv}} : \mathbb{R}^S \times (A^{\mathrm{cv}})^S \times \mathbb{R}_{\geq 0} \to \mathbb{R}^S$,

$$\underline{\mathsf{T}}^{\mathrm{cv}}(v,\pi,\beta) := \mathrm{Proj}_{(A^{\mathrm{cv}})^S}\left(\pi + \beta \tfrac{\partial}{\partial \pi}\underline{\mathsf{F}}^{\mathrm{cv}}(v,\pi)\right),$$

where $\beta \geq 0$ is a learning rate. As a result, we define the *pessimistic value-policy iteration* by

$$v^{k+1} = \underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi^k), \quad \pi^{k+1} = (\underline{\mathsf{T}}^{\mathrm{cv}})^{\ell}(v^{k+1},\pi^k,\beta). \tag{17}$$

In order to analyze the pessimistic value-policy iteration, we introduce the map $\pi^* : \mathbb{R}^S \to (A^{\mathrm{cv}})^S \to (A^{\mathrm{cv}})^S$ by

$$\pi^*(v) = \operatorname{argmax}_{\sigma\in(A^{\mathrm{cv}})^S} \underline{\mathsf{F}}^{\mathrm{cv}}(v,\sigma). \tag{18}$$

The next theorem shows that the interconnection between the pessimistic value iterations and the iteration-distributed optimization described in (17) can be used to approximate the pessimistic optimal value of $\mathcal{IM}^{\mathrm{cv}}$.

*Theorem 4.5 (Value-policy iterations):* Consider the IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$ with an action-space pessimistic relaxation IMDP $\mathcal{IM}^{\mathrm{cv}} = (S, A^{\mathrm{cv}}, [P^{\mathrm{cv}}], [R^{\mathrm{cv}}], \gamma)$ that satisfies Assumption 4.4. Then,

(i) the compact set $\mathcal{X} = \{(v,\pi) \in \mathbb{R}^S_{\geq 0} \times (A^{\mathrm{cv}})^S \mid v \leq \left(\frac{m}{1-\gamma}\right)\mathbb{1}_{|S|}\}$ is a forward invariant set for pessimistic value-policy iterations (17),

(ii) let $\{(v^k,\pi^k)\}_{k=0}^{\infty}$ be the solution of the pessimistic value-policy iteration (17) starting from $(v^0 = 0, \pi^0) \in \mathcal{X}$ with learning rate $\beta = \frac{1}{L}$. Then, for every $k \in \mathbb{Z}_{\geq 1}$,

$$v^k \leq V_{\mathrm{p}}^{\mathrm{cv},*} \leq v^k + \left(\gamma^k \|v^0 - V_{\mathrm{p}}^{\mathrm{cv},*}\|_\infty + \tfrac{(1-\gamma^k)\varepsilon}{1-\gamma}\right)\mathbb{1}_{|S|},$$

where $\varepsilon = \sup_{(v,\pi)\in\mathcal{X}} \left\|\frac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}(v,\pi)\right\|_\infty (1-\frac{c}{L})^{\ell}\|A^{\mathrm{cv}}\|_\infty$.

*Proof:* Regarding part (i), it is easy to show that $\mathcal{X}$ is closed and bounded. So it is compact. Now we assume $v^k \in \mathcal{X}$ and we show that $v^{k+1} \in \mathcal{X}$. Using Proposition 3.3, for every $s \in S$,

$$\begin{aligned} v^{k+1}(s) &= \underline{R}^{\mathrm{cv}}(s,\pi^k(s)) + \gamma\Omega^{\mathcal{IM}^{\mathrm{cv}}}(v^k,s,\pi^k(s)) \\ &\leq \underline{m} + \gamma\sum_{i=1}^{j-1} v^k(i_v)\underline{P}^{\mathrm{cv}}(i_v,s,a) \\ &\quad + \gamma\sum_{i=j+1}^{n} v^k(i_v)\overline{P}^{\mathrm{cv}}(i_v,s,a) + \gamma\xi v^k(j_v), \end{aligned}$$

where $\xi = 1 - \sum_{i=1}^{j-1}\underline{P}(i_v,s,a) - \sum_{i=j+1}^{n}\overline{P}(i_v,s,a)$ and $j = \underline{\iota}^{\mathrm{cv}}(v,s,a)$. Using the fact that $v^k(s) \leq \frac{m}{1-\gamma}$, for every $s \in S$, we get $v^{k+1}(s) \leq \underline{m} + \frac{\gamma m}{1-\gamma} \leq \frac{m}{1-\gamma}$. This implies that $v^{k+1} \leq \frac{m}{1-\gamma}\mathbb{1}_{|S|}$. This means that $v^{k+1} \in \mathcal{X}$ and thus $\mathcal{X}$ is a forward invariant set for the discrete-time system (17).

Regarding part (ii), for every $k \in \mathbb{Z}_{\geq 0}$,

$$v^{k+1} = \underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi^k) \leq \underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi^*(v^k)) = \underline{\mathsf{G}}^{\mathrm{cv}}(v^k).$$

By Theorem 3.2(ii), the map $v \mapsto \underline{\mathsf{G}}^{\mathrm{cv}}(v)$ is monotone and $V_{\mathrm{p}}^{\mathrm{cv},*}$ is the unique fixed point of $v \mapsto \underline{\mathsf{G}}^{\mathrm{cv}}(v)$. Moreover, we have $v_0 = 0 \leq V_{\mathrm{p}}^{\mathrm{cv},*}$. Thus, by the classical monotone comparison theorem [17, Theorem 3.8.1], we have $v^k \leq V_{\mathrm{p}}^{\mathrm{cv},*}$, for every $k \in \mathbb{Z}_{\geq 0}$. We can rewrite the pessimistic value-policy iterations (17) as follows:

$$\begin{aligned} v^{k+1} = \underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi^k) &= \underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi^*(v^k)) \\ &\quad + \int_0^1 \tfrac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}(v^k,\pi^*(v^k)+sy^k)y\,ds. \end{aligned}$$

where $y^k := \pi^k - \pi^*(v^k)$, for every $k \in \mathbb{Z}_{\geq 0}$. Thus,

$$\begin{aligned} v^{k+1} - V_{\mathrm{p}}^{\mathrm{cv},*} &= \underline{\mathsf{G}}^{\mathrm{cv}}(v^k) - \underline{\mathsf{G}}^{\mathrm{cv}}(V_{\mathrm{p}}^{\mathrm{cv},*}) \\ &\quad + \int_0^1 \tfrac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}(v^k,\pi^*(v^k)+sy^k)y\,ds. \end{aligned}$$

By Theorem 3.2(ii), the map $v \mapsto \underline{\mathsf{G}}^{\mathrm{cv}}(v)$ is contracting with respect to the $\ell_\infty$-norm with rate $\gamma$. This implies that

$$\begin{aligned} \|v^{k+1} - V_{\mathrm{p}}^{\mathrm{cv},*}\|_\infty &\leq \gamma\|v^k - V_{\mathrm{p}}^{\mathrm{cv},*}\|_\infty \\ &\quad + \left\|\int_0^1 \tfrac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}(v^k,\pi^*(v^k)+sy^k)y^k\,ds\right\|_\infty. \end{aligned}$$

Now, we bound the second term on the RHS of the above inequality. Using the triangle inequality, for every $k \in \mathbb{Z}_{\geq 0}$,

$$\begin{aligned} &\left\|\int_0^1 \tfrac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}(v^k,\pi^*(v^k)+sy^k)y^k\,ds\right\|_\infty \\ &\quad\leq \int_0^1 \left\|\tfrac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}(v^k,\pi^*(v^k)+sy^k)\right\|_\infty \|y^k\|_\infty ds \\ &\quad\leq \sup_{(v,\pi)\in\mathcal{X}} \left\|\tfrac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}(v,\pi)\right\|_\infty \|y^k\|. \end{aligned}$$

Note that, for every $s \in S$, we have

$$\tfrac{\partial^2}{\partial \pi^2}\underline{\mathsf{F}}^{\mathrm{cv}}(v,\pi)(s) \succeq \tfrac{\partial^2}{\partial \pi^2}\underline{R}(s,\pi(s)) \succeq cI_{|S|}, \tag{19}$$

where the first inequality in equation (19) holds by Proposition 3.3 and the fact that, for every $i \in \{1,\ldots,j\}$, we have $v(i_v) - v(j_v) \geq 0$ and $a \mapsto \underline{P}^{\mathrm{cv}}(i_v,s,a)$ is concave and,

for every $i \in \{j, \dots, n\}$, we have $v(i_v) - v(j_v) \leq 0$ and $a \mapsto \overline{P}^{\mathrm{cv}}(i_v, s, a)$ is convex. The second inequality in equation (19) holds by Assumption 4.4. Therefore, equation (19) implies that the map $\pi \mapsto \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)(s)$ is $c$-strongly concave. Similarly, one can show that $\pi \mapsto \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)(s)$ is $L$-smooth, for every $s \in S$. Therefore,

$$\begin{aligned}\|y^k\| = \|\pi^k - \pi^*(v^k)\|_\infty &\leq \left(1 - \tfrac{c}{L}\right)^\ell \|\pi^{k-1} - \pi^*(v^k)\|_\infty \\ &\leq \left(1 - \tfrac{c}{L}\right)^\ell \|A^{\mathrm{cv}}\|_\infty,\end{aligned}$$

for every $k \in \mathbb{Z}_{\geq 1}$, where the first inequality holds by the fact that $\pi \mapsto \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)(s)$ is $c$-strongly concave and $L$-smooth and using the known results about convergence of projected gradient descent [19, §5.1]. The second inequality holds since $\pi^{k-1}, \pi^*(v^k) \in (A^{\mathrm{cv}})^S$ and $\|(A^{\mathrm{cv}})^S\|_\infty = \|A^{\mathrm{cv}}\|_\infty$. As a result, for every $k \in \mathbb{Z}_{\geq 1}$ $\|v^{k+1} - V_{\mathrm{p}}^{\mathrm{cv},*}\|_\infty \leq \gamma \|v^k - V_{\mathrm{p}}^{\mathrm{cv},*}\|_\infty + \varepsilon$. This means that, for every $k \in \mathbb{Z}_{\geq 0}$

$$\|v^k - V_{\mathrm{p}}^{\mathrm{cv},*}\|_\infty \leq \gamma^k \|v^0 - V_{\mathrm{p}}^{\mathrm{cv},*}\|_\infty + \tfrac{(1-\gamma^k)\varepsilon}{1-\gamma}$$

and, as a result, we get $V_{\mathrm{p}}^{\mathrm{cv},*} \leq v^k + \left(\gamma^k \|v^0 - V_{\mathrm{p}}^{\mathrm{cv},*}\|_\infty + \tfrac{(1-\gamma^k)\varepsilon}{1-\gamma}\right) \mathbb{1}_{|S|}$, for every $k \in \mathbb{Z}_{\geq 1}$. ■

*Example 4.6 (A two-state continuous-action IMDP):* We consider an IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$ with two states $S = \{1, 2\}$ and the continuous action-space $A = [0, 1] \subset \mathbb{R}$ as shown in Figure 1. For every $s, s' \in \{1, 2\}$, and every $a \in [0, 1]$, we define the upper and lower bounds for probability transitions as follows:

$$\underline{P}(s', s, a) = 0.5a, \qquad \overline{P}(s', s, a) = 0.7 + 0.3a.$$

For every $a \in [0, 1]$, we define the lower and the upper bounds for the reward functions as follows:

$$\underline{R}(1, a) = 1 + 4a\sqrt{a} - a^3, \quad \underline{R}(2, a) = 5 - a\sqrt{a}.$$

and we set the discount factor $\gamma = 0.9$. For $\mathcal{IM}$, we consider the IMDP $\mathcal{IM}^{\mathrm{cv}} = (S, A^{\mathrm{cv}}, [P^{\mathrm{cv}}], [R^{\mathrm{cv}}], \gamma)$ with $A^{\mathrm{cv}} = A = [0, 1]$ and with the probability transition bounds $[P^{\mathrm{cv}}] = [P]$. Also the reward bounds are given by

$$\underline{R}^{\mathrm{cv}}(1, a) = 1 + 4a - a^4, \qquad \underline{R}^{\mathrm{cv}}(2, a) = 5 - a^2.$$

The map $a \mapsto \underline{R}^{\mathrm{cv}}(s, a)$ is concave, for every $s \in \{1, 2\}$. Moreover, we have $\underline{R}(s, a) \leq \underline{R}^{\mathrm{cv}}(s, a)$, for every $s \in \{1, 2\}$ and every $a \in [0, 1]$. Thus, $\mathcal{IM}^{\mathrm{cv}}$ is an action-space pessimistic relaxation of $\mathcal{IM}$. We use the distributed optimization implementation of the pessimistic value iterations (17) and Theorem 4.5(ii) with $\beta = 0.01$, $\ell = 1$, and $k = 1000$ iterations to obtain $\left[\begin{smallmatrix}35.2000\\39.2000\end{smallmatrix}\right] \leq V_{\mathrm{p}}^{\mathrm{cv},*} \leq \left[\begin{smallmatrix}52.7000\\56.7000\end{smallmatrix}\right]$. Using the value iterations (15) with the optimization problem (16), one can compute $V_{\mathrm{p}}^{\mathrm{cv},*} = \left[\begin{smallmatrix}43.1820\\43.8891\end{smallmatrix}\right]$.

## V. Conclusions

In this paper, we study IMDPs with continuous action-spaces. We introduce the pessimistic and the optimistic value iterations for IMDPs and show that they are monotone and contracting dynamical systems. Using these observations, we introduce an action-space relaxation of the IMDP and use its value iterations to estimate the optimal policies of the original IMDP. Finally, we propose an iteration-distributed implementation of the value iterations and study its convergence to the optimal values.

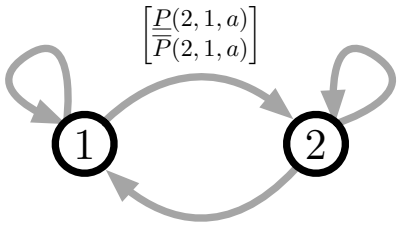


Fig. 1: The state-transition diagram for Example 4.6

## References


[1] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, 2014.

[2] D. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific, 1996.

[3] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[4] G. N. Iyengar, "Robust dynamic programming," *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[5] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[6] S. Li, A. Adje, P.-L. Garoche, and B. Acıkmese, "Bounding fixed points of set-based Bellman operator and Nash equilibria of stochastic games," *Automatica*, vol. 130, p. 109685, 2021.

[7] T. Dick, A. Gyorgy, and C. Szepesvari, "Online learning in Markov decision processes with changing cost sequences," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14, 2014, p. I–512–I–520.

[8] R. Givan, S. Leach, and T. Dean, "Bounded-parameter markov decision processes," *Artificial Intelligence*, vol. 122, no. 1, pp. 71–109, 2000.

[9] G. Delimpaltadakis, M. Lahijanian, M. Mazo Jr, and L. Laurenti, "Interval markov decision processes with continuous action-spaces," *arXiv preprint*, 2022. [Online]. Available: https://arxiv.org/abs/2211.01231

[10] E. M. Wolff, U. Topcu, and R. M. Murray, "Robust control of uncertain Markov decision processes with temporal logic specifications," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 2012, pp. 3372–3379.

[11] J. Jiang, Y. Zhao, and S. Coogan, "Safe learning for uncertainty-aware planning via interval MDP abstraction," *IEEE Control Systems Letters*, vol. 6, pp. 2641–2646, 2022.

[12] M. Lahijanian, S. B. Andersson, and C. Belta, "Temporal logic motion planning and control with probabilistic satisfaction guarantees," *IEEE Transactions on Robotics*, vol. 28, no. 2, pp. 396–409, 2012.

[13] S. Adams, M. Lahijanian, and L. Laurenti, "Formal control synthesis for stochastic neural network dynamic models," *IEEE Control Systems Letters*, vol. 6, pp. 2858–2863, 2022.

[14] M. Dutreix, J. Huh, and S. Coogan, "Abstraction-based synthesis for stochastic systems with omega-regular objectives," *Nonlinear Analysis: Hybrid Systems*, vol. 45, p. 101204, 2022.

[15] S. Haddad and B. Monmege, "Interval iteration algorithm for MDPs and IMDPs," *Theoretical Computer Science*, vol. 735, pp. 111–131, 2018, Reachability Problems 2014: Special Issue.

[16] S. Jafarpour and S. Coogan, "A contracting dynamical system perspective toward interval Markov decision processes," *Preprint*, 2023. [Online]. Available: https://github.com/SaberJafarpour/ContractingIMDP

[17] A. N. Michel, L. Hou, and D. Liu, *Stability of dynamical systems: Continuous, discontinuous, and discrete systems*. Birkhäuser Boston, Inc., Boston, MA, 2008.

[18] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[19] E. K. Ryu and S. Boyd, "Primer on monotone operator methods," *Applied Computational Mathematics*, vol. 15, no. 1, pp. 3–43, 2016.