A Contracting Dynamical System Perspective toward Interval Markov Decision Processes

Saber Jafarpour[1] and Samuel Coogan[1]

***Abstract*— Interval Markov decision processes are a class of uncertain Markov models where the transition probabilities between the states belong to intervals. In this paper, we study the problem of estimating the optimal policies in interval Markov Decision processes (IMDPs) with continuous action-space using the Bellman fixed-point equation. The key element of our analysis is a dynamical system perspective toward the value-iterations in IMDPs. Given an IMDP, we show that the pessimistic and the optimistic value-iterations, i.e., the value-iterations under the assumption of a competitive adversary and cooperative adversary, are monotone dynamical systems and are contracting with respect to the ℓ_∞-norm. Inspired by this dynamical system viewpoint, we introduce another IMDP, called the action-space relaxation IMDP and use it to approximate the optimal policies. 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 pessimistic and optimistic value-iterations can be efficiently solved using tools and techniques from convex analysis. We then turn to the problem of efficient implementation of the pessimistic and the optimistic value-iterations. We consider the policy optimization problems at each step of the value-iterations as a feedback controller for the value function. Using this system-theoretic perspective, we propose a distributed-time optimization implementation of the pessimistic and the optimistic value-iterations. For an action-space relaxation IMDP, we study the performance of the distributed-time implementation of the pessimistic and the optimistic value-iterations using contracting interconnected dynamical systems.**

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 has been proposed to analyze MDPs with parameter uncertainties. In [4], [5], a robust dynamic programming is proposed to study optimal solutions of Markov decision processes with uncertainty in transition probabilities. In [6], a set-value 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 the agent in a MDP interacts with the environment in the presence of an adversary who resolves uncertain transition probabilities [9]. In the literature, IMDPs have been used to analyze a wide range of 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 very 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 or can ruin any guarantee on the optimality of the solutions. Second, it is challenging to implement existing iterative algorithms because they require the solution of an expensive optimization problem at each time-instance, which becomes particularly problematic for large IMDPs or when it is desired to compute optimal policies of the IMDP at runtime. Most of the existing literature on IMDPs focuses on discretizing the action-space and then using known results about discrete-action IMDPs. However, it turns out that this

*This work is partially supported by the National Science Foundation under awards #1749357 and #1931980.

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

approximation is sup-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. 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 competitive adversary and cooperative adversary, respectively. As a minor contribution of this paper, we first present a dynamical system framework to study value-iterations in IMDPs and we study contractivity and monotonicity of the pessimistic and optimistic value-iterations. While the proof of this result is built upon known results in the literature, the new dynamical system perspective provides a framework for the rest of our study in this paper. 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. We consider the pessimistic and the optimistic value-iterations for the action-space relaxation IMDP. As our first main result, we use a dynamical systems perspective to show that these iterations provides over-bounds on the the pessimistic and the optimistic value-iterations of the original IMDP. In particular, the optimal value of these iterations are over-approximation of the the pessimistic and the optimistic optimal values.

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 optimization problem in a time-distributed fashion. We consider the value-iteration and the optimization problem as an interconnected dynamical system and leverage the contractivity of the value-iteration algorithm to provide guarantees for convergence of the interconnected system to the optimal value function.

II. Notations and Mathematical Preliminary

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 $v(1_v) \leq v(2_v) \leq \dots \leq v(n_v)$. 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

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

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 $\mathcal{X} \subseteq \mathbb{R}^n$, we denote the orthogonal projection into $\mathcal{X}$ by $\mathrm{Proj}_{\mathcal{X}}$, i.e., we have $\mathrm{Proj}_{\mathcal{X}}(y) = \mathrm{argmin}_{x \in \mathcal{X}} \|x - y\|_2$, 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 $\pi : A^S$ which assigns an action a to each state s[1]. The space of all policies for $\mathcal{M}$ is denoted by $\Pi^{\mathcal{M}}$. For every policy $\pi \in \Pi^{\mathcal{M}}$, 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 π. The goal is to find a policy $\pi^* \in \Pi^{\mathcal{M}}$ which maximizes the value function $V_\pi^{\mathcal{M}}$, i.e., a policy π^* such that

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

In general, it can be shown that there exists a unique policy $\pi^* \in \Pi^{\mathcal{M}}$ which maximizes the value function $V_\pi^{\mathcal{M}}$ [1, Theorem 6.1.1]. The value function obtained following this optimal policy is denoted by V^*, i.e., $V_{\pi^*}^{\mathcal{M}} = V^*$ and 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* $\mathsf{F} : \mathbb{R}^S \times \Pi^{\mathcal{M}} \to \mathbb{R}^S$ is defined by

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

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 $\mathsf{G}(v) = \max_{\pi \in \Pi^{\mathcal{M}}} \mathsf{F}(v, \pi)$, for every $v \in \mathbb{R}^S$. It is known that the Bellman operator G is contracting with respect to the ℓ_∞-norm and monotone with respect to the standard partial ordering [1, Theorem 6.2.3] and the optimal value function is the fixed point of $v = \mathsf{G}(v)$ [1].

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

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) $[P] \in S \times S \times A \to \text{Interval}_{[0,1]}$ 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 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.

We say that 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 $a \in A$. Policies for IMDPs can be defined similar to the policies for MDPs. A policy for $\mathcal{IM}$ is a vector $\pi \in A^S$ that assigns an action a to each state s. The space of all policies for $\mathcal{IM}$ is denoted by $\Pi^{\mathcal{IM}}$.

Remark 3.1: The following remarks are in order.

(i) *(Comparison with the literature)*: our definition of IMDP generalizes the classical definition in [8], [5] which assumes a finite action-space. This generalization is motivated by, *e.g.,* applications in abstraction of stochastic dynamical systems [14], [9].
(ii) *(Interpretations)*: In the literature, two different interpretations for IMDPs have been proposed. In the first interpretation, an IMDP is considered as an MDP with uncertain parameters [8], [5], whereas in the second interpretation, an IMDP is considered as a MDP interacting with an adversary [6], [9].

Given an IMDP $\mathcal{IM}$ and a policy $\pi \in \Pi^{\mathcal{IM}}$, we would like to 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 $\Delta_{s,a}^{\mathcal{IM}}$ as the set of all $p : S \to [0, 1]$ such that

$$\Delta_{s,a}^{\mathcal{IM}} := \quad \underline{P}(s', s, a) \leq p(s') \leq \overline{P}(s', s, a)$$
$$\mathbb{1}_{|S|}^{\mathsf{T}} p = 1. \tag{5}$$

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

$$\underline{\mathsf{F}}(v, \pi)(s) = \underline{R}(s, \pi(s)) + \gamma \min_{p \in \Delta_{s,\pi(s)}} p^{\mathsf{T}} v,$$
$$\overline{\mathsf{F}}(v, \pi)(s) = \overline{R}(s, \pi(s)) + \gamma \max_{p \in \Delta_{s,\pi(s)}} p^{\mathsf{T}} v, \tag{6}$$

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

$$\underline{\mathsf{G}}(v)(s) = \max_{a \in A} \left\{ \underline{R}(s, a) + \gamma \min_{p \in \Delta_{s,a}} p^{\mathsf{T}} v \right\},$$
$$\overline{\mathsf{G}}(v)(s) = \max_{a \in A} \left\{ \overline{R}(s, a) + \gamma \max_{p \in \Delta_{s,a}} p^{\mathsf{T}} v \right\}. \tag{7}$$

Equivalently, using the vector notation, we have

$$\underline{\mathsf{G}}(v) = \max_{\pi \in \Pi^{\mathcal{M}}} \underline{\mathsf{F}}(v, \pi) \qquad \overline{\mathsf{G}}(v) = \max_{\pi \in \Pi^{\mathcal{M}}} \overline{\mathsf{F}}(v, \pi),$$

for every $v \in \mathbb{R}^S$. The next theorem shows that the interval Bellman (resp. Bellman-policy) operator is monotone and contracting with respect to the ℓ_∞-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 Appendix B.

Theorem 3.2 (Interval Bellman operator): Consider an IMDP $\mathcal{IM}$ with the interval Bellman operator-policy and the interval Bellman operator defined 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 defined in (3) and (4), respectively. Then, the following statements hold:

(i) for every $\pi \in \Pi^{\mathcal{IM}}$, 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 ℓ_∞-norm with rate γ 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 ℓ_∞-norm with rate γ 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 definition 6 requires solving two linear programs in p and can be computationally intractable for large-scale IMDPs. The next theorem provides a closed-from expression for the interval Bellman operator using the lower and upper probability transition bounds. Before we state the next theorem, we 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:

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

where $j = \underline{\iota}(v,s,a)$ and $k = \overline{\iota}(v,s,a)$. In the next proposition, we provide closed-from expressions for the interval Bellman-policy operator. The proof is provided in Appendix A.

Proposition 3.3 (Closed-from of the Bellman 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

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

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

Remark 3.4 (Comparison with the literature):

(i) For finite action space A, Theorem 3.3 recovers the results in [8] and the pseudocode in [8, Figure 8].

(ii) For infinite-dimensional action space A, Theorem 3.3 provides a simpler form compared to [9, Theorem 3.2].

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 a certain type of optimally for the value function. The pessimistic optimal policy $\pi^*_{\mathrm{p}} \in \Pi^{\mathcal{IM}}$ is the unique policy defined by

$$\pi^*_{\mathrm{p}} = \operatorname{argmax}_{\pi\in\Pi^{\mathcal{IM}}}\left(\min_{\mathcal{M}\in\mathcal{IM}} V^{\mathcal{M}}_{\pi}\right),$$

and the pessimistic value function is given by $V^*_{\mathrm{p}} = \min_{\mathcal{M}\in\mathcal{IM}} V^{\mathcal{M}}_{\pi^*_{\mathrm{p}}}$. The optimistic optimal policy $\pi^*_{\mathrm{o}} \in \Pi^{\mathcal{IM}}$ is the unique policy defined by

$$\pi^*_{\mathrm{o}} = \operatorname{argmax}_{\pi\in\Pi^{\mathcal{IM}}}\left(\max_{\mathcal{M}\in\mathcal{IM}} V^{\mathcal{M}}_{\pi}\right),$$

and the optimistic value function is given by $V^*_{\mathrm{o}} = \max_{\mathcal{M}\in\mathcal{IM}} V^{\mathcal{M}}_{\pi^*_{\mathrm{o}}}$. The pessimistic and the optimistic optimal policies have nice interpretations in terms of adversaries. Indeed, the pessimistic optimal policy is the optimal policy in presence of a competitive adversary and the optimistic optimal policy is the presence of a cooperative adversary.

Given an IMDP $\mathcal{IM}$, we define the *pessimistic value-iteration* by

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

and we define the *optimistic value-iteration* by

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

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 Appendix C for the proof.

Theorem 3.5 (value-iterations as dynamical systems): Consider the IMDP $\mathcal{IM} = (S,A,[P],[R],\gamma)$ with the pessimistic and optimistic policies $\pi^*_{\mathrm{p}}, \pi^*_{\mathrm{o}} \in \Pi^{\mathcal{IM}}$ 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^*_{p} and the pessimistic optimal policy π^*_{p} is obtained by

$$\pi^*_{\mathrm{p}} = \operatorname{argmax}_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(V^*_{\mathrm{p}},\pi).$$

(ii) the optimistic value-iteration (12) is a monotone contracting dynamical system with the unique globally exponentially stable equilibrium point V^*_{o} and the optimistic optimal policy π^*_{o} is obtained by

$$\pi^*_{\mathrm{o}} = \operatorname{argmax}_{\pi\in\Pi^{\mathcal{IM}}} \overline{\mathsf{F}}(V^*_{\mathrm{o}},\pi).$$

Remark 3.6 (A dynamical system perspective): The fact that pessimistic and optimistic value-iterations are contracting and the pessimistic and optimistic optimal policies are their fixed points is known in the literature [8, Theorems 10,11,12]. However, Theorem 3.5 provides a discrete-time dynamical system perspective to the pessimistic and optimistic value-iterations (11) and (12) and highlights their less-studied property of monotonicity.

IV. Efficient estimation of optimal policies

Theorem 3.5 provides iterative algorithms for computing the pessimistic and optimistic optimal policies in IMDPs. It turns out that implementing the iterations (11) (resp. iterations (12)) requires solving the following $|S|$ nonlinear optimization problems at each time-step:

$$\underline{\pi}^k = \operatorname{argmax}_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(v^k,\pi) \tag{13}$$

(resp. $\overline{\pi}^k = \operatorname{argmax}_{\pi\in\Pi^{\mathcal{IM}}} \overline{\mathsf{F}}(v^k,\pi)$). This can cause two main challenges for computing the pessimistic and optimistic value-iterations:

(i) in the lack 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 pessimistic and optimistic value-iterations. Thus, it becomes computationally challenging to implement the interval value-iterations for large-scale IMDPs at the runetime.

In order to address the above challenges, we study IMDPs using a dynamical system lens. In particular, in subsection IV-A, we introduce the action-space relaxation IMDP, which replaces the optimization problem (13) with its concave relaxation. In subsection IV-B, we consider the optimization problems (13) as a feedback controller for the pessimistic value-iterations (11). Using this feedback interconnection perspective, we propose to implement pessimistic value-iterations in a time-distributed fashion.

A. Action-space relaxation

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 relaxation): Consider an interval Markov decision process $\mathcal{IM} = (S, A, [P], [R], \gamma)$. An *action-space relaxation* of $\mathcal{IM}$ is an IMDP $\mathcal{IM}^{\mathrm{cv}} = (S, A^{\mathrm{cv}}, [P^{\mathrm{cv}}], [R^{\mathrm{cv}}], \gamma)$ satisfying

(i) we have $A \subseteq A^{\mathrm{cv}}$ and A^{cv} is convex and compact,

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

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

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

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

are concave on A^{cv}.

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

$$a \mapsto \underline{P}^{\mathrm{cv}}(s', s, a)$$

is convex on A^{cv}.

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

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

and we denote the pessimistic optimal value and the pessimistic optimal policy of $\mathcal{IM}^{\mathrm{cv}}$ by $V_{\mathrm{p}}^{\mathrm{cv},*}$ and $\pi_{\mathrm{p}}^{\mathrm{cv},*}$, respectively. Similarly, the optimistic value-iteration of $\mathcal{IM}^{\mathrm{cv}}$ is given by

$$\begin{aligned} v^{k+1} &= \overline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi^k), \\ \pi^k &= \mathrm{argmax}_{\pi \in \Pi^{\mathrm{cv}}} \overline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi), \end{aligned} \tag{15}$$

and we denote the optimistic optimal value and the optimistic optimal policy of $\mathcal{IM}^{\mathrm{cv}}$ by $V_{\mathrm{o}}^{\mathrm{cv},*}$ and $\pi_{\mathrm{o}}^{\mathrm{cv},*}$, respectively.

Given an IMDP $\mathcal{IM}$, the next theorem shows that the Bellman operator of the action-space relaxation $\mathcal{IM}^{\mathrm{cv}}$ is an upper bound for the Bellman operator of $\mathcal{IM}$. Considering the pessimistic and the optimistic value-iterations (14) and (15) as dynamical systems, one can use the classical comparison theorems to show that the pessimistic and the optimistic optimal values of $\mathcal{IM}^{\mathrm{cv}}$ provide over-approximations for the optimal values of $\mathcal{IM}$ and to estimate its pessimistic and optimistic optimal policies.

Theorem 4.2 (Optimal values of action-space relaxation): Consider the IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$ with an associated action-space relaxation IMDP $\mathcal{IM}^{\mathrm{cv}} = (S, A^{\mathrm{cv}}, [P^{\mathrm{cv}}], [R^{\mathrm{cv}}], \gamma)$. The following statements hold:

(i) for every $v \in \mathbb{R}^S$ and every $\pi \in A^S$,

$$\underline{\mathsf{F}}(v, \pi) \leq \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi), \qquad \underline{\mathsf{F}}(v, \pi) \leq \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi).$$

(ii) for every $s \in S$ and every $v \in \mathbb{R}^S$,

$$\pi \to \underline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)(s), \quad \text{and} \quad \pi \to \overline{\mathsf{F}}^{\mathrm{cv}}(v, \pi)(s),$$

are concave functions on A^{cv}.

(iii) for every $v \in \mathbb{R}^S$,

$$\underline{\mathsf{G}}(v) \leq \underline{\mathsf{G}}^{\mathrm{cv}}(v), \qquad \underline{\mathsf{G}}(v) \leq \underline{\mathsf{G}}^{\mathrm{cv}}(v).$$

(iv) we have $V_{\mathrm{p}}^{*} \leq V_{\mathrm{p}}^{\mathrm{cv},*}$ and $V_{\mathrm{o}}^{*} \leq V_{\mathrm{o}}^{\mathrm{cv},*}$.

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 we can compute

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

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 $\overline{P}^{\mathrm{cv}}(i_v, s, a) \geq \overline{P}(i_v, s, a)$ and, for every $i \in \{j, \ldots, n\}$, we have $v(i_v) - v(j_v) \leq 0$ and $\underline{P}^{\mathrm{cv}}(i_v, s, a) \leq \underline{P}(i_v, s, a)$. Given $\pi \in A^S$, we define $p^* \in \mathbb{R}^S$ by

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

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 that $p^* \in \Delta_{s,\pi(s)}^{\mathcal{IM}}$ This implies that

$$\Omega^{\mathcal{IM}^{\mathrm{cv}}}(v, s, \pi(s)) \geq (p^*)^{\mathsf{T}} v \geq \min_{\pi \in \Delta_{s,\pi(s)}} p^{\mathsf{T}} v.$$

As a result, using Theorem 3.3, we get

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

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

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

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. Similarly, one can show that $\pi \mapsto \overline{\mathsf{F}}^{\mathrm{cv}}(v,\pi)(s)$ is a concave function. Regarding part (iii), the fact that $\underline{\mathsf{G}}(v) \leq \underline{\mathsf{G}}^{\mathrm{cv}}(v)$ and $\overline{\mathsf{G}}(v) \leq \overline{\mathsf{G}}^{\mathrm{cv}}(v)$ follows from definition of $\underline{\mathsf{G}}$ and $\overline{\mathsf{G}}$ in (7). Regarding part (iv), by Theorem (3.5)(i), the discrete-time dynamical systems (11) and (14) are monotone and contracting with respect to ℓ_∞-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 [16, Theorem 3.8.1], to conclude that, we have $V_{\mathrm{p}}^* \leq V_{\mathrm{p}}^{\mathrm{cv},*}$. ■

Remark 4.3: The following remarks about Theorem 4.2 are in order.

(i) *(Computational efficiency)*: using the action-space relaxation $\mathcal{IM}^{\mathrm{cv}}$, Theorem 4.2 develops two iterations (14) and (15) for over-approximating the pessimistic and optimistic optimal value functions of the original IMDP $\mathcal{IM}$. Since the interval Bellman-policy operator is concave in π, standard convex optimization algorithms (see [17]) can be employed to solve the optimization problem at each iteration of (14) and (15).

(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 IMDP to approximate the 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. Time-distributed optimization

In practice, estimating the optimal policies using the pessimistic and optimistic value-iterations (14) and (15) requires solving $|S|$ concave optimization problems with m variables at each iteration step, which becomes computationally challenging for IMDPs with large state-space. In this subsection, we consider the pessimistic and the optimistic value-iterations (14) and (15) as feedback interconnected systems. In particular, the pessimistic value-iterations (14) can be considered as the interconnection of a dynamical system described by the pessimistic value-iterations:

$$v^{k+1} = \underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi^k), \tag{16}$$

and a controller described by the optimization problem:

$$\pi^k = \mathrm{argmax}_{\pi\in\Pi^{\mathrm{cv}}} \underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi). \tag{17}$$

Similarly, one can consider the optimistic value-iterations (15) as a feedback interconnected systems. Using this system-theoretic perspective toward interval value-iterations (14) and (15), we propose to implement the feedback controller in a time-distributed fashion. We first need to introduce the following assumption on the IMDP and its action-space relaxation.

Assumption 4.4: For the IMDP $\mathcal{IM} = (S,A,[P],[R],\gamma)$ with an action-space relaxation $\mathcal{IM}^{\mathrm{cv}} = (S,A^{\mathrm{cv}},[P^{\mathrm{cv}}],[R^{\mathrm{cv}}],\gamma)$, the following statements hold:

(i) *(Boundedness of rewards)* there exist $\underline{m},\overline{m} \in \mathbb{R}_{\geq 0}$ such that

$$\sup_{s\in S,a\in A} \overline{R}(s,a) \leq \overline{m}, \qquad \sup_{s\in S,a\in A} \underline{R}(s,a) \leq \underline{m}.$$

(ii) *(Smoothness of relaxation):* the maps

$$a \mapsto \underline{R}^{\mathrm{cv}}(s,a), \qquad a \mapsto \overline{R}^{\mathrm{cv}}(s,a),$$

are continuously differentiable and strongly concave with rate $c \in \mathbb{R}_{\geq 0}$, for every $s',s \in S$, and the maps

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

are continuously differentiable, for every $s',s \in S$.

We also consider the following assumption for implementing the pessimistic and the optimistic value-iterations (14) and (15) of the action-relaxation IMDP $\mathcal{IM}^{\mathrm{cv}}$.

Assumption 4.5 (Optimization algorithm): For the IMDP $\mathcal{IM} = (S,A,[P],[R],\gamma)$ with an action-space relaxation $\mathcal{IM}^{\mathrm{cv}} = (S,A^{\mathrm{cv}},[P^{\mathrm{cv}}],[R^{\mathrm{cv}}],\gamma)$, we use the projected gradient descent algorithm with the constant learning rate β to solve the optimization problem (17). We denote the pessimistic and the optimistic projected gradient descent operators by $\underline{\mathsf{T}}$ and $\overline{\mathsf{T}}$:

$$\begin{aligned}\underline{\mathsf{T}}(v,\pi,\beta) &= \mathrm{Proj}_{A^{\mathrm{cv}}}\left(\pi + \beta\tfrac{\partial}{\partial\pi}\underline{\mathsf{F}}^{\mathrm{cv}}(v,\pi)\right), \\ \overline{\mathsf{T}}(v,\pi,\beta) &= \mathrm{Proj}_{A^{\mathrm{cv}}}\left(\pi + \beta\tfrac{\partial}{\partial\pi}\overline{\mathsf{F}}^{\mathrm{cv}}(v,\pi)\right).\end{aligned}$$

Given $\epsilon \in \mathbb{R}_{\geq 0}$, we define the mixed norm $\|\cdot\|_\epsilon$ on $\mathbb{R}^S \times A^S$ as

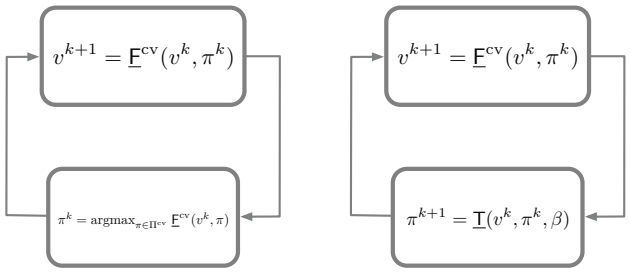
$$\|(v,\pi)\|_\epsilon = \|v\|_\infty + \epsilon\|\pi\|_2. \tag{18}$$

The key idea is to replace the feedback controller described by the optimization problem (17) with one iteration of the pessimistic gradient descent operator $\underline{\mathsf{T}}$, i.e.,

$$\pi^{k+1} = \underline{\mathsf{T}}(v^k,\pi^k,\beta).$$

The interconnection between the pessimistic value-iteration and the pessimistic projected gradient descent operator is shown in Figure 1. As a result, for $\beta \in \mathbb{R}_{\geq 0}$, we can define the *pessimistic value-policy iteration* with learning rate β by

$$\begin{aligned}v^{k+1} &= \underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi^k), \\ \pi^{k+1} &= \underline{\mathsf{T}}(v^k,\pi^k,\beta),\end{aligned} \tag{19}$$



(a) (b)

Fig. 1: Value-iterations as a feedback system: (a) shows the interconnection between the pessimistic value-iterations and the original optimization problem, and (b) shows the feedback interconnection between the pessimistic value-iterations and the pessimistic projected gradient descent operator $\underline{\mathsf{T}}$.

and we can define the *optimistic value-policy iteration* with learning rate β by

$$\begin{aligned} v^{k+1} &= \overline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi^k), \\ \pi^{k+1} &= \overline{\mathsf{T}}(v^k, \pi^k, \beta). \end{aligned} \tag{20}$$

In the next theorem, we show that for small enough learning rate, the interconnection between the pessimistic (resp. optimistic) value-iterations and the time-distributed optimization described in (19) (resp. in (20)) converges to the pessimistic (resp. optimistic) optimal value of $\mathcal{IM}^{\mathrm{cv}}$.

Theorem 4.6 (Value-policy iterations): Suppose that the IMDP $\mathcal{IM} = (S, A, [P], [R], \gamma)$ satisfies the Assumption 4.4 and 4.5. Then the following statements hold:

(i) the compact set

$$\mathcal{X} = \{(v, \pi) \in \mathbb{R}^S_{\geq 0} \times (A^{\mathrm{cv}})^S \mid v \leq \left(\tfrac{m}{1-\gamma}\right) \mathbb{1}_{|S|}\}$$

is a forward invariant set for pessimistic value-policy iterations (19).

(ii) for every $\eta < \min\{\frac{1-\gamma}{M}, \left(\frac{c}{N}\right)^{\frac{1}{3}}, \sqrt{c}\}$, where

$$\begin{aligned} N &= \sqrt{|S|} \sup_{(v,\pi)\in\mathcal{X}} \left\| \tfrac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial v \partial \pi}(v, \pi) \right\|_2, \\ M &= \sup_{(v,\pi)\in\mathcal{X}} \left\| \tfrac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}(v, \pi) \right\|_\infty, \end{aligned}$$

the pessimistic value-policy iterations (19) with the learning rate $\beta = \eta^2$ is contracting on $\mathcal{X}$ with respect to the mixed norm (18) with $\epsilon = \frac{1}{\eta}$.

(iii) the pessimistic value-policy iterations (19) converges to $(V_{\mathrm{p}}^{\mathrm{cv},*}, \pi_{\mathrm{p}}^{\mathrm{cv},*})$ satisfying $V_{\mathrm{p}}^* \leq V_{\mathrm{p}}^{\mathrm{cv},*}$.

Moreover, the following statements hold:

(iv) the compact set

$$\mathcal{Y} = \{(v, \pi) \in \mathbb{R}^S_{\geq 0} \times (A^{\mathrm{cv}})^S \mid v \leq \left(\tfrac{m}{1-\gamma}\right) \mathbb{1}_{|S|}\}$$

is a forward invariant set for the optimistic value-policy iterations (20).

(v) for every $\eta < \min\{\frac{1-\gamma}{M}, \left(\frac{c}{N}\right)^{\frac{1}{3}}, \sqrt{c}\}$, where

$$\begin{aligned} N &= \sqrt{|S|} \sup_{(v,\pi)\in\mathcal{X}} \left\| \tfrac{\partial \overline{\mathsf{F}}^{\mathrm{cv}}}{\partial v \partial \pi}(v, \pi) \right\|_2, \\ M &= \sup_{(v,\pi)\in\mathcal{X}} \left\| \tfrac{\partial \overline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}(v, \pi) \right\|_\infty, \end{aligned}$$

the optimistic value-policy iterations (19) with learning rate $\beta = \eta^2$ is contracting on $\mathcal{Y}$ with respect to the mixed norm (18) with $\epsilon = \frac{1}{\eta}$.

(vi) the optimistic value-policy iterations (20) converges to $(V_{\mathrm{o}}^{\mathrm{cv},*}, \pi_{\mathrm{o}}^{\mathrm{cv},*})$ satisfying $V_{\mathrm{o}}^* \leq V_{\mathrm{o}}^{\mathrm{cv},*}$.

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 Theorem 3.3, for every $s \in S$,

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

where $\xi = 1 - \sum_{i=1}^{j-1} \overline{P}(i_v, s, a) - \sum_{i=j+1}^{n} \underline{P}(i_v, s, a)$ 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 (19).

Regarding part (ii), we consider $k \mapsto (v^k, \pi^k)$ and $k \mapsto (w^k, \sigma^k)$ as two trajectories of the discrete-time system (19). Thus, for every $k \in \mathbb{Z}_{\geq 0}$, we have

$$\begin{aligned} &\|(v^{k+1}, \pi^{k+1}) - (w^{k+1}, \sigma^{k+1})\|_\epsilon \\ &= \|v^{k+1} - w^{k+1}\|_\infty + \epsilon \|\pi^{k+1} - \sigma^{k+1}\|_2. \end{aligned}$$

Moreover, we can compute

$$\begin{aligned} \|v^{k+1} - w^{k+1}\|_\infty &= \|\underline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi^k) - \underline{\mathsf{F}}^{\mathrm{cv}}(w^k, \sigma^k)\|_\infty, \\ \|\pi^{k+1} - \sigma^{k+1}\|_2 &= \left\|\underline{\mathsf{T}}(v^k, \pi^k) - \underline{\mathsf{T}}(w^k, \sigma^k)\right\|_2. \end{aligned}$$

Note that, using triangle inequality, we have

$$\begin{aligned} &\left\|\underline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi^k) - \underline{\mathsf{F}}^{\mathrm{cv}}(w^k, \sigma^k)\right\|_\infty \\ &\quad \leq \left\|\underline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi^k) - \underline{\mathsf{F}}^{\mathrm{cv}}(w^k, \sigma^k)\right\|_\infty \\ &\qquad + \|\underline{\mathsf{F}}^{\mathrm{cv}}(w^k, \pi^k) - \underline{\mathsf{F}}^{\mathrm{cv}}(w^k, \sigma^k)\|. \end{aligned}$$

Moreover, by Theorem 3.2(i), we have $\left\|\underline{\mathsf{F}}^{\mathrm{cv}}(v^k, \pi^k) - \underline{\mathsf{F}}^{\mathrm{cv}}(w^k, \sigma^k)\right\|_\infty \leq \lambda \|v^k - w^k\|_\infty$. Also,

using the Mean value Inequality [18, Proposition 2.4.7],

$$\begin{aligned}\big\|\underline{\mathsf{F}}^{\mathrm{cv}}(w^k,\pi^k)-\underline{\mathsf{F}}^{\mathrm{cv}}(w^k,\sigma^k)\big\| &\le \sup_{(v,\pi)\in\mathcal{X}}\left\|\frac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}\right\|_\infty \|\pi^k-\sigma^k\|_\infty \\ &\le \sup_{(v,\pi)\in\mathcal{X}}\left\|\frac{\partial \underline{\mathsf{F}}^{\mathrm{cv}}}{\partial \pi}\right\|_\infty \|\pi^k-\sigma^k\|_2,\end{aligned}$$

where the last inequality holds using $\|x\|_\infty \le \|x\|_2$, for every $x\in\mathbb{R}^n$. This implies that

$$\begin{aligned}\big\|\underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi^k)-\underline{\mathsf{F}}^{\mathrm{cv}}(w^k,\sigma^k)\big\|_\infty \le \lambda\|v^k-w^k\|_\infty \\ +\eta M\|\pi^k-\sigma^k\|_2.\end{aligned}$$

Moreover, using triangle inequality,

$$\begin{aligned}\|\underline{\mathsf{T}}(v^k,\pi^k)-\underline{\mathsf{T}}(w^k,\sigma^k)\|_2 \le \|\underline{\mathsf{T}}(w^k,\pi^k)-\underline{\mathsf{T}}(w^k,\sigma^k)\|_2 \\ +\|\underline{\mathsf{T}}(v^k,\pi^k)-\underline{\mathsf{T}}(w^k,\pi^k)\|_2. \qquad (21)\end{aligned}$$

Regarding the first term in the RHS of inequality (21),

$$\begin{aligned}&\|\underline{\mathsf{T}}(w^k,\pi^k)-\underline{\mathsf{T}}(w^k,\sigma^k)\|_2 \\ &\le \big\|\pi^k+\beta\tfrac{\partial}{\partial\pi}\underline{\mathsf{F}}^{\mathrm{cv}}(w^k,\pi^k)-\sigma^k-\beta\tfrac{\partial}{\partial\pi}\underline{\mathsf{F}}^{\mathrm{cv}}(w^k,\pi^k)\big\|_2 \\ &\le \sup_{(v,\pi)\in\mathcal{X}}\left\|I_n+\beta\tfrac{\partial^2\underline{\mathsf{F}}^{\mathrm{cv}}}{\partial\pi^2}\right\|_2\|\sigma^k-\pi^k\|_2,\end{aligned}$$

where the first inequality holds because the projection operator is non-expansive with respect to the ℓ_2-norm and the second inequality holds by the Mean Value Inequality [18, Proposition 2.4.7]. Moreover, by Assumption 4.4, the maps $a\mapsto \underline{R}^{\mathrm{cv}}(s,a)$ are strongly concave with rate $c\in\mathbb{R}_{\ge 0}$, for every $s\in S$. This means that the map $\pi\mapsto \underline{\mathsf{F}}(v,\pi)(s)$ is strongly concave with rate $c\in\mathbb{R}_{\ge 0}$, for every $v\in\mathbb{R}^S$. Note that, by definition of η, we have $\beta=\eta^2<c$ and therefore, by [19, Section 5.1], we get

$$\sup_{(v,\pi)\in\mathcal{X}}\left\|I_n+\beta\tfrac{\partial^2\underline{\mathsf{F}}^{\mathrm{cv}}}{\partial\pi^2}\right\|_2\le 1-\beta c.$$

Regarding the second term in the RHS of inequality (21),

$$\begin{aligned}&\|\underline{\mathsf{T}}(v^k,\pi^k)-\underline{\mathsf{T}}(w^k,\pi^k)\|_2 \\ &\le \beta\,\big\|\tfrac{\partial}{\partial\pi}\underline{\mathsf{F}}^{\mathrm{cv}}(v^k,\pi^k)-\tfrac{\partial}{\partial\pi}\underline{\mathsf{F}}^{\mathrm{cv}}(w^k,\pi^k)\big\|_2 \\ &\le \beta\sup_{(v,\pi)\in\mathcal{X}}\left\|\tfrac{\partial^2\underline{\mathsf{F}}^{\mathrm{cv}}}{\partial v\partial\pi}\right\|_2\|v^k-w^k\|_2 \\ &\le \beta\sqrt{|S|}\sup_{(v,\pi)\in\mathcal{X}}\left\|\tfrac{\partial^2\underline{\mathsf{F}}^{\mathrm{cv}}}{\partial v\partial\pi}\right\|_2\|v^k-w^k\|_\infty,\end{aligned}$$

where the first inequality holds because the projection operator is non-expansive with respect to the ℓ_2-norm, the second inequality holds by the Mean Value Inequality [18, Proposition 2.4.7], and the third inequality holds because $\|x\|_2\le\sqrt{n}\|x\|_\infty$, for every $x\in\mathbb{R}^n$. This implies that

$$\begin{aligned}\|\underline{\mathsf{T}}(v^k,\pi^k)-\underline{\mathsf{T}}(w^k,\sigma^k)\|_2 \le (1-\beta c)\|\pi^k-\sigma^k\|_2^2 \\ +\beta N\|v^k-w^k\|_\infty.\end{aligned}$$

This means that, for every $k\in\mathbb{Z}_{\ge 0}$, we have

$$\begin{bmatrix}\|v^{k+1}-w^{k+1}\|_\infty\\ \frac{1}{\eta}\|\pi^{k+1}-\sigma^{k+1}\|_2\end{bmatrix}\le\begin{bmatrix}\gamma & \eta N\\ \frac{\beta}{\eta}M & (1-\beta c)\end{bmatrix}\begin{bmatrix}\|v^k-w^k\|_\infty\\ \frac{1}{\eta}\|\pi^k-\sigma^k\|_2\end{bmatrix}$$

We pick $\beta=\eta^2$ and we can easily check that, if $\eta<\min\{\frac{1-\gamma}{M},\big(\frac{c}{N}\big)^{\frac{1}{3}}\}$, then we have $\left\|\begin{bmatrix}\gamma & \eta N\\ \frac{\beta}{\eta}M & (1-\beta c)\end{bmatrix}\right\|_1=\xi<1$. This implies that, for every $\eta<\min\{\frac{1-\gamma}{M},\big(\frac{c}{N}\big)^{\frac{1}{3}}\}$, we have

$$\begin{aligned}\|v^{k+1}-w^{k+1}\|_\infty+\tfrac{1}{\eta}\|\pi^{k+1}-\sigma^{k+1}\|_2 \\ \le \xi(\|v^k-w^k\|_\infty+\tfrac{1}{\eta}\|\pi^k-\sigma^k\|_2).\end{aligned}$$

Thus the discrete-time dynamical system (19) is contracting with respect to the mixed norm (18) with $\epsilon=\frac{1}{\eta}$.

Regarding part (iii), it is easy to see that $(V_{\mathrm{p}}^{\mathrm{cv},*},\pi_{\mathrm{p}}^{\mathrm{cv},*})$ is an equilibrium point of the discrete-time dynamical system (19). Thus, by part (ii), every trajectory of the system (19) converges to $(V_{\mathrm{p}}^{\mathrm{cv},*},\pi_{\mathrm{p}}^{\mathrm{cv},*})$. The fact that $V_{\mathrm{p}}^*\le V_{\mathrm{p}}^{\mathrm{cv},*}$ follows from Theorem 4.2(iv). The proofs of parts (iv), (v), and (vi) are similar and we omit them. ■

Example 4.7 (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 2. 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:

$$\begin{aligned}\underline{R}(1,a)&=1+3a-a^3, & \underline{R}(2,a)&=5-\sin(a),\\ \overline{R}(1,a)&=1+4\sqrt{a}-a^2, & \overline{R}(2,a)&=5-a^2.\end{aligned}$$

and we set the discount factor $\gamma=0.9$. For $\mathcal{IM}$, we consider the the action-space relaxation $\mathcal{IM}^{\mathrm{cv}}=(S,A^{\mathrm{cv}},[P^{\mathrm{cv}}],[R^{\mathrm{cv}}],\gamma)$ with $A^{\mathrm{cv}}=A=[0,1]$ and with the probability transition bounds

$$\begin{aligned}\underline{P}^{\mathrm{cv}}(s',s,a)&=\underline{P}(s',s,a)=0.5a,\\ \overline{P}^{\mathrm{cv}}(s',s,a)&=\overline{P}(s',s,a)=0.7+0.3a,\end{aligned}$$

for every $s,s'\in S$ and every $a\in[0,1]$. Also the reward bounds are given by

$$\begin{aligned}\underline{R}^{\mathrm{cv}}(1,a)&=1+3a-a^4, & \underline{R}^{\mathrm{cv}}(2,a)&=5,\\ \overline{R}^{\mathrm{cv}}(1,a)&=1+4a-a^2, & \overline{R}^{\mathrm{cv}}(2,a)&=5-a^2,\end{aligned}$$

It is easy to check that $a\mapsto\underline{R}^{\mathrm{cv}}(s,a)$ and $a\mapsto\overline{R}^{\mathrm{cv}}(s,a)$ are concave, for every $s\in\{1,2\}$. Moreover, we have $\underline{R}(s,a)\le\underline{R}^{\mathrm{cv}}(s,a)$ and $\overline{R}(s,a)\le\overline{R}^{\mathrm{cv}}(s,a)$, for every $s\in\{1,2\}$ and every $a\in[0,1]$. Thus, $\mathcal{IM}^{\mathrm{cv}}$ is a valid action-space relaxation of $\mathcal{IM}$. We use the distributed optimization implementation of the the pessimistic and the optimistic value iterations (shown in in (19) and 19) to compute the pessimistic and the optimistic optimal policies. The learning rate is $\beta=0.01$ and satisfies the conditions in Theorem 4.6. Using this time-distributed implementation, we compute The optimistic optimal value as $V_{\mathrm{o}}^*=\begin{bmatrix}47.1686\\ 47.9651\end{bmatrix}$ and the optimistic optimal policy given by $\pi_{\mathrm{o}}^*=\begin{bmatrix}1.0000\\ 0.1075\end{bmatrix}$. We also compute the pessimistic optimal value is given by

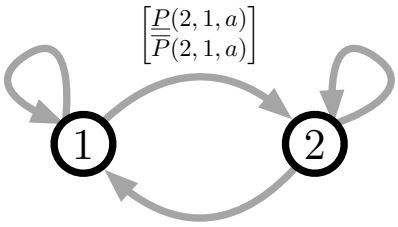


Fig. 2: The state-transition diagram for the interval Markov decision process $\mathcal{IM}$ given in Example 4.7

$V_{\mathrm{p}}^* = \begin{bmatrix}39\\ 41\end{bmatrix}$ with the pessimistic optimal policy given by $\pi_{\mathrm{p}}^* = \begin{bmatrix}1.0000\\ 0\end{bmatrix}$.

V. Conclusions

In this paper, we focus on IMDP with continuous action-spaces and propose computationally efficient methods for approximating their optimal polices. We show that the pessimistic and the optimistic value-iterations of IMDPs can be considered as monotone and contracting dynamical systems. Using these observations, we introduce a relaxation of the IMDP with computationally efficient pessimistic and optimistic value-iterations. We then use this relaxation to estimate the pessimistic and optimistic policies of the original IMDP. We also propose a time-distributed implementation of the pessimistic and the optimistic value-iterations for large-scale IMDPs or when the runtime computation of the optimal policies are desirable.

Appendices

A. Proof of Proposition 3.3

We start by showing that $\max_{p\in\Delta_{s,a}^{\mathcal{IM}}} p^{\mathsf{T}}v = \Omega^{\mathcal{IM}}(v,s,a)$. In the course of this proof we set $j = \underline{\iota}(v,s,a)$. Note that, we can compute

$$\begin{aligned} p^{\mathsf{T}}v &= \sum_{i=1}^{j}(v(i_v)-v(j_v))p(i_v)\\ &\quad+\sum_{i=j}^{n}(v(i_v)-v(j_v))p(i_v)+v(j_v),\end{aligned}$$

where the equality holds using the fact that $\mathbb{1}^{\mathsf{T}}p = 1$. The above equality implies that

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

where the second inequality holds because $\underline{P}(i_v,s,a) \leq p(i_v) \leq \overline{P}(i_v,s,a)$ and because $v(i_v)-v(j_v)\geq 0$ if $i\leq j$ and $v(i_v)-v(j_v)\leq 0$ if $i>j$. This means that, for every $p\in\Delta_{s,a}^{\mathcal{IM}}$, we have $p^{\mathsf{T}}v \geq \Omega^{\mathcal{IM}}(v,s,a)$. Now we show that $\min_{p\in\Delta_{s,a}^{\mathcal{IM}}} p^{\mathsf{T}}v = \Omega^{\mathcal{IM}}(v,s,a)$. Let $v\in\mathbb{R}^n$ and define $p^*\in\mathbb{R}^S$ as follows:

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

where $\xi = 1-\sum_{i=1}^{j-1}\overline{P}(i_v,s,a)-\sum_{i=j+1}^{n}\underline{P}(i_v,s,a)$. Note that by definition of $\underline{\iota}(v,s,a)$, we have $p^*\leq \mathbb{1}_{|S|}$. Moreover $\mathbb{1}_{|S|}^{\mathsf{T}}p^* = 1$ and $\underline{P}(s',s,a)\leq p^*(s')\leq \overline{P}(s',s,a)$, for every $s'\in S$. Therefore $p^*\in\Delta_{s,a}^{\mathcal{IM}}$ and with this choice of p^*, we have $(p^*)^{\mathsf{T}}v = \Omega^{\mathcal{IM}}(v,s,a)$. The proof of $\max_{p\in\Delta_{s,a}^{\mathcal{IM}}} p^{\mathsf{T}}v = \Lambda(v,s,a)$ is similar and we omit it for the sake of brevity.

B. Proof of Theorem 3.2

Regarding part (i), consider $v,w\in\mathbb{R}^S$ such that $v\leq w$. As a result, we have $p^{\mathsf{T}}v\leq p^{\mathsf{T}}w$, for every $p\in\Delta_{s,a}^{\mathcal{IM}}$ and every $s,a\in S\times A$. Therefore, for every $s\in S$ and every $\pi\in A^S$,

$$\begin{aligned}\underline{\mathsf{F}}(v,\pi)(s) &= \underline{R}(s,\pi(s))+\min_{p\in\Delta_{s,\pi(s)}^{\mathcal{IM}}} p^{\mathsf{T}}v\\ &\leq \underline{R}(s,\pi(s))+\min_{p\in\Delta_{s,\pi(s)}^{\mathcal{IM}}} p^{\mathsf{T}}w = \underline{\mathsf{F}}(w,\pi)(s)\end{aligned}$$

This implies that the operator $v\mapsto\underline{\mathsf{F}}(v,a)$ is monotone. Moreover, for every $\pi\in A^{|S|}$,

$$\begin{aligned}&\|\underline{\mathsf{F}}(v,\pi)-\underline{\mathsf{F}}(w,\pi)\|_\infty\\ &\qquad=\gamma\max_{s\in S}\left|\min_{p\in\Delta_{s,\pi(s)}^{\mathcal{IM}}} p^{\mathsf{T}}v-\min_{p\in\Delta_{s,\pi(s)}^{\mathcal{IM}}} p^{\mathsf{T}}w\right|\end{aligned}$$

Let $s\in S$ and note that $\min_{p\in\Delta_{s,\pi(s)}^{\mathcal{IM}}} p^{\mathsf{T}}v \leq \min_{p\in\Delta_{s,\pi(s)}^{\mathcal{IM}}} p^{\mathsf{T}}w$. Since $\Delta_{s,\pi(s)}^{\mathcal{IM}}$ is compact, there exists $p^*\in\Delta_{s,\pi(s)}^{\mathcal{IM}}$ such that $\min_{p\in\Delta_{s,\pi(s)}^{\mathcal{IM}}} p^{\mathsf{T}}v = (p^*)^{\mathsf{T}}v$. This

implies that, for every $s \in S$,

$$\begin{aligned}\Big|\min_{p\in\Delta^{\mathcal{IM}}_{s,\pi(s)}} p^{\mathsf{T}}v - \min_{p\in\Delta^{\mathcal{IM}}_{s,\pi(s)}} p^{\mathsf{T}}w\Big| \\ = \min_{p\in\Delta^{\mathcal{IM}}_{s,\pi(s)}} p^{\mathsf{T}}w - \min_{p\in\Delta^{\mathcal{IM}}_{s,\pi(s)}} p^{\mathsf{T}}v \\ = \min_{p\in\Delta^{\mathcal{IM}}_{s,\pi(s)}} p^{\mathsf{T}}w - (p^*)^{\mathsf{T}}v \le (p^*)^{\mathsf{T}}(v-w) \le \|v-w\|_\infty.\end{aligned}$$

As a result, we get $\|\underline{\mathsf{F}}(v,\pi) - \underline{\mathsf{F}}(v,\pi)\|_\infty \le \gamma\|v-w\|_\infty$. This means that $v \mapsto \underline{\mathsf{F}}(v,\pi)$ is contracting with respect to the ℓ_∞-norm with the rate γ. Similarly, one can show that $v \mapsto \overline{\mathsf{F}}(v,\pi)$ is monotone and contracting with respect to the ℓ_∞-norm with the rate γ. To show the inequalities in (8), note that, for every $a \in A$ and every $s, s' \in S$, we have $\underline{P}(s',s,a) \le P(s',s,a) \le \overline{P}(s',s,a)$. Using the definition of $\Delta^{\mathcal{IM}}_{s,a}$ in equation (5), we get that

$$\min_{p\in\Delta^{\mathcal{IM}}_{s,a}} p^{\mathsf{T}}v \le \sum_{s'\in S} P(s',s,a)v(s') \le \max_{p\in\Delta^{\mathcal{IM}}_{s,a}} p^{\mathsf{T}}v$$

for every $s \in S$ and every $a \in A$. Thus, for every $v \in \mathbb{R}^S$ and $\pi \in A^S$, we have

$$\begin{aligned}\underline{\mathsf{F}}(v,\pi)(s) &= \underline{R}(s,\pi(s)) + \gamma \min_{p\in\Delta^{\mathcal{IM}}_{s,\pi(s)}} p^{\mathsf{T}}v \\ &\le R(s,\pi(s)) + \gamma \sum_{s'\in S} P(s',s,\pi(s))v(s') = \mathsf{F}(v,\pi)(s) \\ &\le \overline{R}(s,\pi(s)) + \gamma \max_{p\in\Delta^{\mathcal{IM}}_{s,\pi(s)}} p^{\mathsf{T}}v = \overline{\mathsf{F}}(v,\pi)(s).\end{aligned}$$

Regarding part (ii), consider $v, w \in \mathbb{R}^S$ such that $v \le w$. Then, using part (i), we have $\underline{\mathsf{F}}(v,\pi) \le \underline{\mathsf{F}}(w,\pi)$, for every $\pi \in \Pi^{\mathcal{IM}}$. This implies that

$$\underline{\mathsf{G}}(v) = \max_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(v,\pi) \le \max_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(w,\pi) = \underline{\mathsf{G}}(w).$$

This means that $\underline{\mathsf{G}}$ is monotone. On the other hand, let $\pi^* \in \Pi^{\mathcal{IM}}$ be such that $\mathsf{F}(v,\pi^*) = \max_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(v,\pi)$. Therefore, we have

$$\begin{aligned}\|\underline{\mathsf{G}}(v) - \underline{\mathsf{G}}(w)\|_\infty &= \|\max_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(v,\pi) - \max_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(w,\pi)\|_\infty \\ &= \max_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(w,\pi) - \max_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(v,\pi) \\ &= \max_{\pi\in\Pi^{\mathcal{IM}}} \underline{\mathsf{F}}(w,\pi) - \underline{\mathsf{F}}(v,\pi^*) \\ &\le \underline{\mathsf{F}}(w,\pi^*) - \underline{\mathsf{F}}(v,\pi^*) \le \gamma\|v-w\|_\infty,\end{aligned}$$

where the last inequality holds by contractivity of $v \mapsto \underline{\mathsf{F}}(v,\pi^*)$ proved in part (i). Thus, $\underline{\mathsf{G}}$ is contracting with respect to the ℓ_∞-norm with the rate γ. Similarly, one can show that $\overline{\mathsf{G}}$ is monotone and contracting with respect to the ℓ_∞-norm with the rate γ. Finally the inequalities 9 follows from inequalities (8) in part (i) and definition of $\underline{\mathsf{G}}$ and $\overline{\mathsf{G}}$.

C. Proof of Theorem 3.5

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. [Online]. Available: http://www.jstor.org/stable/25151652

[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. D., A. G., and C. Szepesvári, "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] A. N. Michel, L. Hou, and D. Liu, *Stability of dynamical systems: Continuous, discontinuous, and discrete systems*, ser. Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 2008.

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

[18] R. Abraham, J. E. Marsden, and T. S. Ratiu, *Manifolds, Tensor Analysis, and Applications*, 2nd ed., ser. Applied Mathematical Sciences, 1988, vol. 75.

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