Set-based value operators for non-stationary and uncertain Markov decision processes

Automatica 171 (2025) 111970

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Set-based value operators for non-stationary and uncertain Markov decision processes☆

Sarah H.Q. Li [a,*], Assalé Adjé [c], Pierre-Loïc Garoche [d], Behçet Açıkmeşe [b]

[a] Automatic Control Laboratory, ETH Zürich, Zürich, Switzerland
[b] Department of Aeronautics and Astronautics, University of Washington, Seattle, USA
[c] LAMPS, Université de Perpignan Via Domitia, Perpignan, France
[d] École Nationale de l'Aviation Civile, Université de Toulouse, Toulouse, France

ARTICLE INFO

Article history:
Received 27 July 2022
Received in revised form 5 September 2023
Accepted 6 September 2024
Available online xxxx

Keywords:
Markov decision process
Contraction operator
Stochastic control
Decision making
Autonomy

ABSTRACT

This paper analyzes finite-state Markov Decision Processes (MDPs) with nonstationary and uncertain parameters via set-based fixed point theory. Given compact parameter ambiguity sets, we demonstrate that a family of contraction operators, including the Bellman operator and the policy evaluation operator, can be extended to set-based contraction operators with a unique fixed point—a compact value function set. For non-stationary MDPs, we show that while the value function trajectory diverges, its Hausdorff distance from this fixed point converges to zero. In parameter uncertain MDPs, the fixed point's extremum value functions are equivalent to the min–max value function in robust dynamic programming under the rectangularity condition. Furthermore, we show that the rectangularity condition is a sufficient condition for the fixed point to contain its own extremum value functions. Finally, we derive novel guarantees for probabilistic path planning in capricious wind fields and stratospheric station-keeping.

1. Introduction

Markov decision process (MDP) is a versatile model for decision making in stochastic environments and is widely used in trajectory planning (Al-Sabban, Gonzalez, & Smith, 2013), robotics (Van Hoof, Hermans, Neumann, & Peters, 2015), and operations research (Doshi, Goodwin, Akkiraju, & Verma, 2005). Given state–action costs and transition probabilities, finding an optimal policy of the MDP is equivalent to solving for the fixed point *value function* of the corresponding Bellman operator. However, finding the fixed point value function may not be straight forward when the MDP model does not accurately describe its operation environment. In MDP applications such as traffic light control and dexterous manipulation, the MDP model is prone to two sources of inaccuracy: *environmental non-stationarity* and *parameter uncertainty*.

Different from an MDP's internal stochasticity, environmental non-stationarity refers to the time-varying changes in the MDP parameters, in particular as induced by external factors or the presence of interfering decision makers in an unpredictable way. Standard dynamic programming under environmental non-stationarity has no convergence guarantees and can be shown to diverge.

Example 1 (*Navigating in Non-Stationary Wind*)**.** An autonomous aircraft navigates a non-stationary wind field to reach a non-stationary target. The non-stationary wind field alternates between N known wind patterns over time and is predictable for the next time step but unpredictable in the long term. This creates environmental non-stationarity in both the aircraft's transition dynamics as well as the target position. Should its policy optimization use the average wind pattern, the worst-case wind pattern, or the short-term wind forecast?

On the other hand, parameter uncertainty refers to the discrepancy between the modeling parameters used in computation vs. the parameters that accurately model a time-invariant system. Minimizing the risk or worst-case failure for a parameter-uncertain MDP can be tackled from the min–max approach via robust MDPs (Iyengar, 2005; Mannor, Mebel, & Xu, 2016) and distributionally robust MDPs (Yang, 2017) under rectangularity assumptions on the parameter uncertainty's state–action space structure (Iyengar, 2005; Mannor et al., 2016).

☆ This research is partly funded by National Science Foundation grant CMMI-210563 and the University of Washington Aero&Astro Condit fellowship. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Subhrakanti Dey under the direction of Editor Florian Dorfler.

* Corresponding author.
E-mail addresses: sarahli@control.ee.ethz.ch (S.H.Q. Li), assale.adje@univ-perp.fr (A. Adjé), Pierre-Loic.Garoche@enac.fr (P.-L. Garoche), behcet@uw.edu (B. Açıkmeşe).

https://doi.org/10.1016/j.automatica.2024.111970

In this paper, we develop a set-theoretic extension to the Bellman operator that is applicable to MDPs experiencing both environmental non-stationarity and parameter uncertainty. We use this approach to derive convergence guarantees for environmentally non-stationary MDPs under the Hausdorff distance and to extend the rectangularity assumptions for parameter-uncertain MDPs.

Contributions. We propose the set-extensions of *value operators*: a general class of contraction operators that includes the Bellman operator and the policy evaluation operator. For these value operators, we extend their input–output domain to sets of value functions, show the existence of a compact *fixed point* and prove that value iteration over sets of value functions converges. Under non-stationarity assumptions, we show that the Hausdorff distance between the value iteration trajectory and the fixed point of the value operator always converges to zero. Under MDP parameter uncertainty, we show that the rectangularity assumption in the min–max MDP model is sufficient for the fixed point of the value operator to contain its own extremal value functions. Given a value operator and a compact parameter uncertainty set, we present an algorithm that computes the bounds of the corresponding fixed point set. Finally, we apply our results to the wind-assisted navigation of high altitude platform systems relevant to space exploration (Wolf et al., 2010) and show that our algorithms can be used to derive policies with better guarantees.

Related research. MDP with parameter uncertainty is well-studied in robust control and reinforcement learning. In control theory, the worst-case cost-to-go with respect to state-decoupled parameter uncertainties is derived via a minmax variation of the Bellman operator in Givan, Leach, and Dean (2000), Iyengar (2005), Nilim and El Ghaoui (2005) and Wiesemann, Kuhn, and Rustem (2013). The cost-to-go under parameter uncertainty with coupling between states and time steps is similarly bounded in Goyal and Grand-Clement (2022) and Mannor et al. (2016). The effect of statistical uncertainty on the optimal cost-to-go is studied in Mannor et al. (2016), Nilim and El Ghaoui (2005), Wiesemann et al. (2013) and Yang (2017). Recently, parameter-uncertain MDPs gained traction in the reinforcement learning community due to the presence of uncertainty in real world problems such as traffic signal control and multi-agent coordination (Kumar, Zhou, Tucker, & Levine, 2020; Lecarpentier & Rachelson, 2019; Padakandla, KJ, & Bhatnagar, 2020). Most RL research extends the worst-case analysis to methods such as Q-learning and SARSA. Recently, methods for value-based RL using non-contracting operators have been investigated in Bellemare, Ostrovski, Guez, Thomas, and Munos (2016).

As an alternative to optimizing for the worst-case, distributionally robust MDP optimizes a risk metric over the value functions of an MDP whose parameters are described by a known set of probability distributions. Distributionally robust MDPs boil down to solving a min–max MDP formulation (Mannor et al., 2016; Xu & Mannor, 2010; Yang, 2017; Yu & Xu, 2015) over an ambiguity set, and require a rectangularity condition to be solvable.

We do not optimize the worst-case cost-to-go by assuming adversarial MDP parameter selection. Instead, we derive a set of cost-to-go values that is invariant with respect to a compact parameter uncertainty set. We continue from our previous work (Li, Adjé, Garoche, & Açıkmeşe, 2021), in which we analyzed the set-based Bellman operator for cost uncertainty only.

Notation: A set of N elements is given by $[N] = \{0, \dots, N-1\}$. We denote the set of matrices of i rows and j columns with real (non-negative) entries as $\mathbb{R}^{i\times j}$ ($\mathbb{R}_+^{i\times j}$), respectively. Matrices and some integers are denoted by capital letters, X, while sets are denoted by cursive typeset $\mathcal{X}$. The set of all *compact subsets* of $\mathbb{R}^d$ is denoted by $\mathcal{K}(\mathbb{R}^d)$. The column vector of ones of size $N \in \mathbb{N}$ is denoted by $\mathbf{1}_N = [1, \dots, 1]^T \in \mathbb{R}^{N\times 1}$. The identity matrix of size S is denoted by I_S. The simplex of dimension S is denoted by

$$\Delta_S = \{p \in \mathbb{R}^S \mid \mathbf{1}_S^\top p = 1,\ p \geqslant 0\}. \tag{1}$$

A vector $h \in \mathbb{R}^S$ has equivalent notation $(h_1, \dots, h_s)$, where h_s is the value of h in the sth coordinate, $s \in [S]$.

2. Discounted infinite-horizon MDP

A *discounted infinite-horizon finite state MDP* is given by $([S], [A], P, C, \gamma)$, where $\gamma \in (0, 1)$ is a discount factor, $[S] = \{1, \dots, S\}$ is a finite set of states and $[A] = \{1, \dots, A\}$ is a finite set of actions. Without loss of generality, assume that every action is admissible from every state.

MDP Costs. $C \in \mathbb{R}^{S\times A}$ is the matrix encoding the MDP cost. Each $C_{sa} \in \mathbb{R}$ is the cost of taking action $a \in [A]$ from state $s \in [S]$. We also denote the cost of all actions at state s by $c_s = [C_{s1}, \dots, C_{sA}] \in \mathbb{R}^A$, such that $C = [c_1, \dots, c_S]^\top$.

MDP Transition Dynamics. The transition probabilities when action a is taken from state s are given by $p_{sa} \in \Delta_S$. Collectively, all possible transition probabilities from state $s \in [S]$ are given by the matrix $P_s = [p_{s1}, \dots, p_{sA}] \in \Delta_S^A \subset \mathbb{R}^{S\times A}$, and all possible transition probabilities in the MDP are given by the matrix $P = [P_1, \dots, P_S] \in \Delta_S^{SA} \subset \mathbb{R}^{S\times SA}$.

MDP Policy. The policy is denoted as $\pi = [\pi_1, \dots, \pi_S] \in \Delta_A^S$, where the ath element of $\pi_s \in \Delta_A$ is the conditional probability of action a being chosen from state s. We also use $\pi(s)$ to denote the corresponding discrete random variable with probability density π_s.

MDP Objective. Under policy π, the decision maker's expected cost-to-go is given per state by

$$V_s^\pi \coloneqq \mathbb{E}_s\left\{\sum_{t=0}^{\infty} \gamma^t C_{s^t a^t} \mid s^0 = s, a^t \sim \pi(s^t)\right\},\ \forall\, s \in [S], \tag{2}$$

where $\mathbb{E}_s\{\cdot\}$ is the expected value of the input with respect to initial state s, and (s^t, a^t) are the state and action at time t. The decision maker's objective is to minimize V_s for all $s \in [S]$. We denote the minimum as $V_s^\star$.

$$V_s^\star \coloneqq \min_{\pi\in\Delta_A^S} \mathbb{E}_s\left\{\sum_{t=0}^{\infty} \gamma^t C_{s^t a^t} \mid s^0 = s, a^t \sim \pi(s^t)\right\}, \tag{3}$$

Under policy π_s, the expected immediate cost at s is given by $c_s^\top \pi_s \in \mathbb{R}$ and the expected transition probabilities from s is given by $P_s\pi_s \in \Delta_S$.

Remark 1. Although *value function* is the standard term for the expected cost-to-go, we use value vector in this paper to emphasize that the cost-to-go values of finite MDPs belong in a finite-dimensional space.

2.1. Value operators

We can find the optimal policy by finding the value vector that minimizes the objective (3). Typical solution methods utilize *order preserving* (Schröder, 2003, Def. 3.1), α*-contractive operators* whose fixed points are the MDP's optimal value vectors (e.g. Bellman operator (Puterman, 2014, Thm. 6.2.3), Q-value operator (Melo, 2001)).

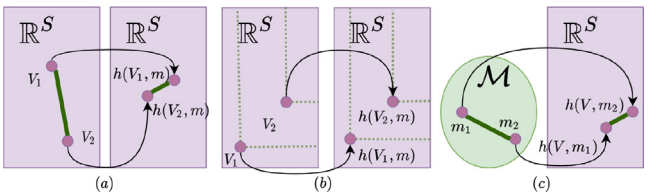


Fig. 1. Visualization of the three value operator properties. (a) α-contraction on $\mathbb{R}^S$, (b) Order preservation on $\mathbb{R}^S$, and (c) $K(V)$-Lipschitz in input space $\mathcal{M}$.

Definition 1 (*α-Contraction*)**.** Let $(\mathcal{X}, d)$ be a metric space with metric d. The operator $H : \mathcal{X} \mapsto \mathcal{X}$ is an α-contraction if and only if there exists $\alpha \in [0, 1)$ such that

$$d\big(H(V), H(V')\big) \leqslant \alpha d(V, V'), \quad \forall\, V,\ V' \in \mathcal{X}. \tag{4}$$

Definition 2 (*Order Preservation*)**.** Let $(\mathcal{X}, \leqslant)$ be an ordered space with partial order $\leqslant$. The operator $H : \mathcal{X} \mapsto \mathcal{X}$ is order preserving if for all $V, V' \in \mathcal{X}$ such that $V \leqslant V'$, $H(V) \leqslant H(V')$.

These operators are typically locally Lipschitz in the MDP parameter space.

Definition 3 (*$K(V)$-Lipschitz*)**.** Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space with metric $d_{\mathcal{X}}$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be a metric space with metric $d_{\mathcal{Y}}$. The operator $H : \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{X}$ is *$K(V)$-Lipschitz* with respect to $\mathcal{M} \subset \mathcal{Y}$ if for all $V \in \mathcal{X}$, there exists $K(V) \in \mathbb{R}_+$ such that

$$d_{\mathcal{X}}\big(H(V, m), H(V, m')\big) \leqslant K(V) d_{\mathcal{Y}}(m, m'), \ \forall m, m' \in \mathcal{M}. \tag{5}$$

Remark 2. The α-contraction property is a special case of Lipschitz continuity when the input and output spaces are identical and the Lipschitz constant is less than 1.

To capture operators with these properties, we define a value operator that takes inputs: the value vector, the cost, and the transition dynamics. The cost and transition dynamics are selected from a parameter set $\mathcal{M}$.

Definition 4 (*Value Operator*)**.** Consider the operator h, given by

$$h : \mathbb{R}^S \times \mathcal{M} \mapsto \mathbb{R}^S, \ \ \mathcal{M} \subseteq \mathbb{R}^{S \times A} \times \Delta_S^{SA}. \tag{6}$$

We say that h (6) is a **value operator** (see Fig. 1) on $\mathbb{R}^S \times \mathcal{M}$ if

(1) For all $m \in \mathcal{M}$, $h(\cdot, m)$ is α-contractive on $\mathbb{R}^S$.
(2) For all $m \in \mathcal{M}$, $h(\cdot, m)$ is order preserving in $\mathbb{R}^S$.
(3) For all $V \in \mathbb{R}^S$, $h(V, m)$ is $K(V)$-Lipschitz on $\mathcal{M}$.

Remark 3. Definition 4 and the subsequent results can be extended to the space of Q-value functions by swapping $\mathbb{R}^S$ for $\mathbb{R}^{SA}$ in Definition 4 (Melo, 2001).

An immediate consequence an operator being α-contractive and order-preserving on $\mathbb{R}^S$ is that it is continuous on $\mathbb{R}^S \times \mathcal{M}$.

Lemma 1 (*Continuity*)**.** *If h (6) is a value operator on $\mathbb{R}^S \times \mathcal{M}$, h is continuous on $\mathbb{R}^S \times \mathcal{M}$.*

See Appendix B for proof. Examples of value operators include the Bellman operator and the policy evaluation operators when the cost and transition dynamics are input parameters rather than fixed parameters.

Definition 5 (*Policy Evaluation Operator*)**.** Given a policy $\pi \in \Delta_A^S$, the vector-valued operator $g^\pi = (g_1^\pi, \ldots, g_S^\pi) : \mathbb{R}^S \times \mathbb{R}^{S \times A} \times \Delta_S^{SA} \mapsto \mathbb{R}^S$ is defined per state as

$$g_s^\pi(V, C, P) \coloneqq c_s^\top \pi_s + \gamma (P_s \pi_s)^\top V, \ \ \forall s \in [S]. \tag{7}$$

Given (C, P), $g^\pi(\cdot, C, P) : \mathbb{R}^S \mapsto \mathbb{R}^S$ is a vector-valued operator whose fixed point is the expected cost-to-go of the MDP $([S], [A], C, P, \gamma)$ under π, denoted as $V^\pi(C, P)$ (Puterman, 2014, Thm. 6.2.5).

$$V^\pi(C, P) = g^\pi(V^\pi, C, P), \ \ V^\pi(C, P) \in \mathbb{R}^S. \tag{8}$$

When the context is clear, we denote $V^\pi(C, P)$ as V^π.

Definition 6 (*Bellman Operator*)**.** The vector-valued operator $f = (f_1, \ldots, f_S) : \mathbb{R}^S \times \mathbb{R}^{S \times A} \times \Delta_S^{SA} \mapsto \mathbb{R}^S$ is defined per each state as

$$f_s(V, C, P) \coloneqq \inf_{\pi_s \in \Delta_A} g_s^\pi(V, C, P), \ \ \forall s \in [S]. \tag{9}$$

The corresponding optimal policy $\pi^\star = (\pi_1^\star, \ldots, \pi_s^\star)$ is defined per state as $\pi_s^\star \in \operatorname{argmin}_{\pi_s \in \Delta_A} g_s^\pi(V, C, P)$ (9). One such policy is defined for all $(s, a) \in [S] \times [A]$ by

$$\pi_{sa}^\star \coloneqq \begin{cases} > 0 & a \in \operatorname*{argmin}_{a \in [A]} C_{sa} + \gamma p_{sa}^\top V \\ 0 & \text{otherwise,} \end{cases} \qquad \sum_{a \in [A]} \pi_{sa}^\star = 1. \tag{10}$$

where $\operatorname{argmin}_{a \in [A]}(h)$ is the set of minimizing actions for the function h. An optimal policy in the form (10) always exists for a discounted infinite horizon MDP (Puterman, 2014, Thm 6.2.10). For a parameters (C, P), $f(\cdot, C, P) : \mathbb{R}^S \mapsto \mathbb{R}^S$ is a vector operator whose fixed point is the optimal cost-to-go for the MDP $([S], [A], P, C, \gamma)$, denoted as $V^B(C, P)$.

$$V^B(C, P) = f(V^B, C, P), \ \ V^B(C, P) \in \mathbb{R}^S. \tag{11}$$

When the context is clear, we denote $V^B(C, P)$ as V^B.

We show that both (7) and (9) are value operators.

Lemma 2. *The Bellman operator (9) and the policy evaluation operators (7) for all $\pi \in \Delta_A^S$ are value operators on $\mathbb{R}^S \times \mathcal{M}$ where $\mathcal{M} \subseteq \mathbb{R}^{S \times A} \times \Delta_S^{SA}$ (6).*

See Appendix C for proof.

Remark 4. Beyond the policy evaluation operator and the Bellman operator, many algorithms in reinforcement learning can be cast as value operators. For example, the Q-learning operator (Melo, 2001) and the off-policy temporal difference operator (Chen, Maguluri, Shakkottai, & Shanmugam, 2021) are both value operators on $\mathbb{R}^{SA}$.

3. Set-based value operators

We now define set-based value operators with respect to a compact set of MDP parameters, beginning with Hausdorff-type set distances.

Definition 7 (*Point-to-Set Distance*)**.** The distance between a value vector and a set $\mathcal{V} \subseteq \mathbb{R}^S$ is given by

$$W \mapsto d(W, \mathcal{V}) \coloneqq \inf_{V \in \mathcal{V}} \|W - V\|_\infty. \tag{12}$$

On the space of compact subsets of $\mathbb{R}^S$, given by $\mathcal{K}(\mathbb{R}^S)$, the distance between value vector sets extends (12) and is given by the Hausdorff distance (Henrikson, 1999).

Definition 8 (*Set-to-Set Distance*)**.** The Hausdorff distance between two compact value vector sets $\mathcal{V}, \mathcal{W} \subseteq \mathbb{R}^S$ is given by

$$d_{\mathcal{K}}(\mathcal{V}, \mathcal{W}) \coloneqq \max \left\{ \sup_{V \in \mathcal{V}} d(V, \mathcal{W}), \ \sup_{W \in \mathcal{W}} d(W, \mathcal{V}) \right\}. \tag{13}$$

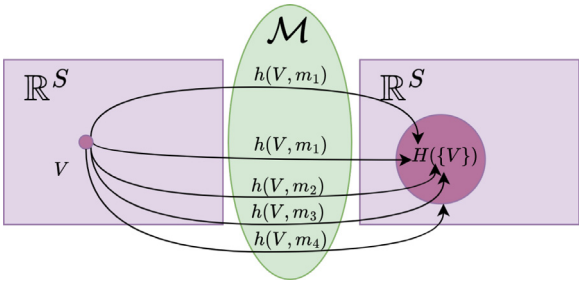


Fig. 2. Visualization of the set-based operator $H(\mathcal{V})$ applied to a singleton set $\mathcal{V} = \{V\} \subset \mathbb{R}^S$: $H(\mathcal{V}) = \cup_{m\in\mathcal{M}} h(V, m)$.

We use $\big(\mathcal{K}(\mathbb{R}^S), d_{\mathcal{K}}\big)$ to denote the metric space formed by the set of all compact subsets of $\mathbb{R}^S$ under the Hausdorff distance $d_{\mathcal{K}}$. The induced Hausdorff space is complete if and only if the original metric space is complete (Henrikson, 1999, Thm 3.3). Therefore, $\big(\mathcal{K}(\mathbb{R}^S), d_{\mathcal{K}}\big)$ is a complete metric space.

Given a value operator h (6), we ask: what is the *set* of possible value vectors when the MDP exhibits environmental non-stationarity captured by $\mathcal{M}$? To resolve this, we define the set-based value operator H.

Definition 9 (*Set-Based Value Operator*)**.** The set-valued operator H is induced by h on $\mathbb{R}^S \times \mathcal{M}$ (6) and is defined as

$$H(\mathcal{V}) \coloneqq \{h(V, m) \mid (V, m) \in \mathcal{V} \times \mathcal{M}\} \subseteq \mathbb{R}^S, \tag{14}$$

where $\mathcal{V} \subseteq \mathbb{R}^S$ is a subset of the value vector space (see Fig. 2).

We denote the set-based value operator induced by the Bellman operator (9) and the policy evaluation operators (7) as F and G^π, respectively, such that for any value vector set $\mathcal{V} \subseteq \mathbb{R}^S$,

$$F(\mathcal{V}) \coloneqq \{f(V, C, P) \mid (V, C, P) \in \mathcal{V} \times \mathcal{M}\}, \tag{15}$$

$$G^\pi(\mathcal{V}) \coloneqq \{g^\pi(V, C, P) \mid (V, C, P) \in \mathcal{V} \times \mathcal{M}\}, \forall\, \pi \in \Delta_A^S. \tag{16}$$

The set-based Bellman F is the union over all one-step *optimal* value vectors, which may result from different policies, while G^π is the union over all value vectors from the same policy π.

We ask the following: is there a set of value vectors that is invariant with respect to H? Similar to the value operators h from Definition 4, we affirmatively answer this by demonstrating that H is α-contractive on $\mathcal{K}(\mathbb{R}^S)$.

Theorem 1. *If h is a value operator on $\mathbb{R}^S \times \mathcal{M}$ (6) and $\mathcal{M}$ is compact, then the induced set value operator H (14) satisfies*

(1) For all $\mathcal{V} \in \mathcal{K}(\mathbb{R}^S)$, $H(\mathcal{V}) \in \mathcal{K}(\mathbb{R}^S)$;
(2) H is α-contractive on $\big(\mathcal{K}(\mathbb{R}^S), d_{\mathcal{K}}\big)$ (13) with a unique fixed point set $\mathcal{V}^\star$ given by

$$H(\mathcal{V}^\star) = \mathcal{V}^\star, \quad \mathcal{V}^\star \in \mathcal{K}(\mathbb{R}^S); \tag{17}$$

(3) The sequence $\{\mathcal{V}^k\}_{k\in\mathbb{N}}$ where $\mathcal{V}^{k+1} = H(\mathcal{V}^k)$ converges to $\mathcal{V}^\star$ for any $\mathcal{V}^0 \in \mathcal{K}(\mathbb{R}^S)$.

In particular, these hold for F (15) and G^π (16), whose fixed point sets are denoted as $\mathcal{V}^B$ and $\mathcal{V}^\pi$, respectively.

$$F(\mathcal{V}^B) = \mathcal{V}^B \in \mathcal{K}(\mathbb{R}^S),\ G^\pi(\mathcal{V}^\pi) = \mathcal{V}^\pi \in \mathcal{K}(\mathbb{R}^S),\ \forall \pi \in \Delta_A^S. \tag{18}$$

Proof. The first statement follows from Lemma 1, since the image of a compact set by a continuous function is compact (Rudin et al., 1964). Let us prove the second statement: for some $\beta \in (0, 1)$, for all, $\mathcal{V}, \mathcal{V}' \in \mathcal{K}(\mathbb{R}^S)$:

$$\begin{aligned} &d_{\mathcal{K}}\big(H(\mathcal{V}), H(\mathcal{V}')\big) \\ &= \max\left\{ \sup_{\substack{V\in\mathcal{V}\\ m\in\mathcal{M}}} d\big(h(V, m), H(\mathcal{V}')\big),\ \sup_{\substack{V'\in\mathcal{V}'\\ m'\in\mathcal{M}}} d\big(h(V', m'), H(\mathcal{V})\big) \right\} \\ &\leqslant \beta d_{\mathcal{K}}(\mathcal{V}, \mathcal{V}') \end{aligned}$$

Take $(V, m) \in \mathcal{V} \times \mathcal{M}$, then $d\big(h(V, m), H(\mathcal{V}')\big) \leqslant \inf_{V'\in\mathcal{V}'} \|h(V, m) - h(V', m)\|_\infty \leqslant \alpha \inf_{V'\in\mathcal{V}'} \|V - V'\|_\infty$ holds from the α-contractive property of h. Finally,

$$\begin{aligned} \sup_{\substack{V\in\mathcal{V}\\ m\in\mathcal{M}}} d\big(h(V, m), H(\mathcal{V}')\big) &\leqslant \alpha \sup_{V\in\mathcal{V}} \inf_{V'\in\mathcal{V}'} \|V - V'\|_\infty \\ &\leqslant \alpha d_{\mathcal{K}}(\mathcal{V}, \mathcal{V}') \end{aligned}$$

We use the same technique to prove that

$$\sup_{\substack{V'\in\mathcal{V}'\\ m'\in\mathcal{M}}} d\big(h(V', m'), H(\mathcal{V})\big) \leqslant \alpha d_{\mathcal{K}}(\mathcal{V}, \mathcal{V}'). \tag{19}$$

Finally, $d_{\mathcal{K}}\big(H(\mathcal{V}), H(\mathcal{V}')\big) \leqslant \alpha d_{\mathcal{K}}(\mathcal{V}, \mathcal{V}')$. From the Banach fixed point theorem and the completeness of $\big(\mathcal{K}(\mathbb{R}^S), d_{\mathcal{K}}\big)$ (Henrikson, 1999, Thm 3.3), H has a unique fixed point $\mathcal{V}^\star$ in $\mathcal{K}(\mathbb{R}^S)$.

The third point is a consequence of the Banach fixed point theorem. Finally, (18) holds because f and g^π are value operators (6) on $\mathbb{R}^S \times \mathcal{M}$. □

Remark 5. Theorem 1's results can be extended to continuous state–action domains if the operator h satisfies Definition 4.

A consequence of Theorem 1 is the existence of a *set-based value iteration*, given by

$$\mathcal{V}^{k+1} = H(\mathcal{V}^k),\ \mathcal{V}^0 \in \mathcal{K}(\mathbb{R}^S). \tag{20}$$

Analogous to the standard value iteration, (20) defines a sequence of value vector sets in $\mathcal{K}(\mathbb{R}^S)$ that converges to the fixed point set $\mathcal{V}^\star \in \mathcal{K}(\mathbb{R}^S)$. In the next section, we demonstrate how this set-based approach leads to convergence results for non-stationary value iteration.

4. Properties of the fixed point set

We analyze the properties of the fixed point set $\mathcal{V}^\star$ under non-stationary value iteration in this section and derive contraction operators that bound $\mathcal{V}^\star$.

4.1. Non-stationary value iteration

Given a value operator h on $\mathbb{R}^S \times \mathcal{M}$, we consider value iteration under a dynamic parameter uncertainty model, as discussed in Nilim and El Ghaoui (2005), where at every iteration, a new set of MDP parameters m^k is chosen from $\mathcal{M}$ such that

$$V^{k+1} = h(V^k, m^k),\ V^0 \in \mathbb{R}^S,\ m^k \in \mathcal{M}, \forall k \in \mathbb{N}. \tag{21}$$

In the robust MDP approach (Iyengar, 2005; Nilim & El Ghaoui, 2005), m^k is adversarially modified such that (21). We consider a different scenario in which m^k is chosen from the closed and bounded set $\mathcal{M}$. In this scenario, convergence of V^k in $\mathbb{R}^S$ will not occur for all possible sequences of $\{m^k\}_{k\in\mathbb{N}}$. However, we can show convergence results on the set domain by leveraging our fixed point analysis of the set-based operator H (14).

Proposition 1. *Let $\mathcal{V}^\star$ be the fixed point set of the set-based value operator H (14) induced by h on $\mathbb{R}^S \times \mathcal{M}$ (6). If the non-stationary value iteration (21) satisfies $\{m^k\}_{k\in\mathbb{N}} \subset \mathcal{M}$, then the sequence $\{V^k\}_{k\in\mathbb{N}}$ defined by (21) satisfies*

(1) $\lim_{k\to+\infty} d(V^k, \mathcal{V}^\star) = 0$,
(2) *there exists a sub-sequence* $\{V^{\varphi(k)}\}_{k\in\mathbb{N}}$ *that converges to a point in* $\mathcal{V}^\star$ *as* $\lim_{k\to\infty} V^{\varphi(k)} \in \mathcal{V}^\star$.

Proof. Let $\{\mathcal{V}^k\}_{k\in\mathbb{N}}$ be a set sequence defined by $\mathcal{V}^0 = \{V^0\}$ and $\mathcal{V}^{k+1} = H(\mathcal{V}^k)$, where H (14) is the set operator induced by h on $\mathbb{R}^S \times \mathcal{M}$. We first show statement (1). From Theorem 1, $\lim_{k\to\infty} \mathcal{V}^k$ converges to $\mathcal{V}^\star$ in $d_{\mathcal{K}}$. Therefore, $0 \leqslant d(V^k, \mathcal{V}^\star) = \inf_{y\in\mathcal{V}^\star} \|V^k - y\|_\infty \leqslant \sup_{x\in\mathcal{V}^k} \inf_{y\in\mathcal{V}^\star} \|x - y\|_\infty \leqslant d_{\mathcal{K}}(\mathcal{V}^k, \mathcal{V}^\star) \to 0$ as k tends to $+\infty$.

Next, for all $k \in \mathbb{N}$, there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$, $d(V^n, \mathcal{V}^\star) \leqslant (k+1)^{-1}$. We define the strictly increasing function $\psi_1 : \mathbb{N} \to \mathbb{N}$, such that $\psi_1(0) = 0$ and for all $k \neq 0$, $\psi_1(k) \coloneqq \min\{N > \psi_1(k-1) : \forall n \geqslant N,\ d(V^n, \mathcal{V}^\star) < (k+1)^{-1}\}$. Then, for all $k \in \{1, 2, \ldots\}$, there exists $y^{\psi_1(k)} \in \mathcal{V}^\star$ such that $\|V^{\psi_1(k)} - y^{\psi_1(k)}\|_\infty < (k+1)^{-1}$. As $\mathcal{V}^\star$ is compact, there exists $\psi_2 : \mathbb{N} \to \mathbb{N}$ strictly increasing such that $(y^{\psi_1(\psi_2(k))})_k$ converges to some $y^\star \in \mathcal{V}^\star$ (Rudin et al., 1964, Thm 3.6). Finally, let $\varepsilon > 0$, there exist $K_1, K_2 \in \mathbb{N}$ such that for all $l \geqslant K_1$, $(\psi_2(l))^{-1} < \varepsilon/2$ and for all $l' \geqslant K_2$, $\|y^{\psi_1(\psi_2(l'))} - y^\star\|_\infty < \varepsilon/2$. So, taking $k \geqslant \max\{K_1, K_2\}$, we have $\|V^{\psi_1(\psi_2(k))} - y^\star\|_\infty \leqslant \|V^{\psi_1(\psi_2(k))} - y^{\psi_1(\psi_2(k))}\|_\infty + \|y^{\psi_1(\psi_2(k))} - y^\star\|_\infty \leqslant \varepsilon$ and $(V^{\psi_1(\psi_2(k))})_k$ converges to $y^\star \in \mathcal{V}^\star$. □

Beyond bounding the asymptotic behavior of value vector trajectories under non-stationary parameters, the fixed point set $\mathcal{V}$ also contains all fixed points of the value operator $h(\cdot, m)$ when $m \in \mathcal{M}$ (6) is fixed.

Corollary 1. *Let* h (6) *be a value operator on* $\mathbb{R}^S \times \mathcal{M}$ *where* $\mathcal{M}$ *is compact. For all* $m \in \mathcal{M}$, *if* $V = h(V, m) \in \mathbb{R}^S$ *and* $\mathcal{V}^\star$ *is the fixed point set of the induced set-based value operator* H (14), $V \in \mathcal{V}^\star$.

Proof. We construct sequence $\{V^k\}$ where $V^{k+1} = h(V^k, m)$ and $V^0 = V$. Then $V^k = V$ for all $k \in \mathbb{N}$. From the second point of Proposition 1, $V \in \mathcal{V}^\star$ follows. □

Furthermore, we can bound the transient behavior of (21) when V^0 is an element of the fixed point set $\mathcal{V}^\star$.

Corollary 2 (*Transient Behavior*)**.** *Let* $\mathcal{V}^\star$ *be the fixed point of the set-based value operator* H (14) *induced by* h *on* $\mathbb{R}^S \times \mathcal{M}$. *If* $\mathcal{M}$ *is compact and* $V^0 \in \mathcal{V}^\star$, *then the sequence generated by* (21) *satisfies* $\{V^k\}_{k\in\mathbb{N}} \subseteq \mathcal{V}^\star$.

Proof. As the fixed point of H (14), $\mathcal{V}^\star$ (17) satisfies $\mathcal{V}^\star = H(\mathcal{V}^\star)$, then the following is true by definition of H: if $V^k \in \mathcal{V}^\star$, then $V^{k+1} = h(V^k, m^k) \in \mathcal{V}^\star$. If $V^0 \in \mathcal{V}^\star$, then $\{V^k\}_{k\in\mathbb{N}} \subseteq \mathcal{V}^\star$ follows by induction. □

Proposition 1 and Corollary 2 bound the asymptotic and transient behavior of the sequence $\{h(V^k, m^k)\}_{k\in\mathbb{N}}$ from (21), irrespective of the convergence of the value vector sequence. This is a more general result than the classic convergence results for MDPs and robust MDPs.

4.2. Bounds of the fixed point set

Since $\mathcal{V}^\star$ is compact, it must have finite supremum and infimum value vectors. We show that it is possible to define contraction operators whose fixed points correspond to these extremal elements.

Greatest and least elements. We define the supremum and infimum elements of a value vector set $\mathcal{V} \in \mathcal{K}(\mathbb{R}^S)$ element-wise as follows,

$$\overline{V}_s \coloneqq \sup_{V\in\mathcal{V}} V_s,\ \underline{V}_s \coloneqq \inf_{V\in\mathcal{V}} V_s, \forall\, s \in [S]. \tag{22}$$

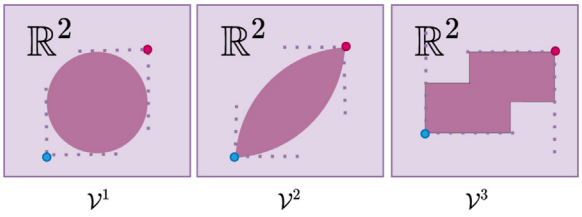


Fig. 3. The greatest least bounds of three different value function sets $\mathcal{V}^i \in \mathbb{R}^2$, where $(0, 0)$ the origin is located on the lower left. Sets $\mathcal{V}^2$ and $\mathcal{V}^3$ contain their own greatest and least elements, but $\mathcal{V}^1$ does not.

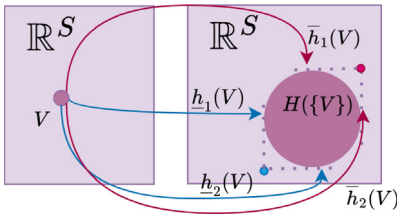


Fig. 4. We visualize $\underline{h}/\overline{h}$ for $H(\mathcal{V})$ on $\mathbb{R}^S \times \mathcal{M}$. The input $\mathcal{V} = \{V\} \in \mathbb{R}^2$. Because $\underline{h}_1$ and $\underline{h}_2$ are achieved for two different $m \in \mathcal{M}$, the resulting $\underline{h}(V)$ lies outside of $H(\mathcal{V})$.

If a set $\mathcal{V} \subseteq \mathbb{R}^S$ is compact, the coordinate-wise supremum and infimum values are achieved by elements of $\mathcal{V}$. However, it is not obvious that a single element in $\mathcal{V}$ can achieve the minimum over all states—i.e., $\overline{V}(\underline{V})$ may not be an element of $\mathcal{V}$. This is illustrated in Fig. 3. We introduce the following bound operators (see Fig. 4).

Definition 10 (*Bound Operators*)**.** The bound operators induced by the value operator h on $\mathbb{R}^S \times \mathcal{M}$ are coordinate-wise defined at each $s \in [S]$ as

$$\underline{h}_s(V) = \inf_{m\in\mathcal{M}} h_s(V, m),\ \overline{h}_s(V) = \sup_{m\in\mathcal{M}} h_s(V, m). \tag{23}$$

Our goal is to bound the fixed point set $\mathcal{V}$ of the set-based value operator H (14) using the bound operators $\underline{h}/\overline{h}$ (23). First we show that $\underline{h}/\overline{h}$ are α-contractive and order preserving on $\mathbb{R}^S$.

Lemma 3 (α-*Contraction*)**.** *If* h (6) *is a value operator on* $\mathbb{R}^S \times \mathcal{M}$ *and* $\mathcal{M}$ *is compact, then* $\underline{h}$ *and* $\overline{h}$ (23) *are* α*-contractions with fixed points* $\underline{X}, \overline{X}$, *respectively.*

$$\overline{h}(\overline{X}) = \overline{X},\quad \underline{h}(\underline{X}) = \underline{X},\ \underline{X}, \overline{X} \in \mathbb{R}^S. \tag{24}$$

Proof. From Lemma 1, h is continuous and $\mathcal{M}$ is compact, then for all $X, Y \in \mathbb{R}^S$, there exists $\hat{m}(s) \in \mathcal{M}$ such that $\underline{h}_s(Y) = h_s\big(Y, \hat{m}(s)\big)$ and $\underline{h}_s(X) \leqslant h_s\big(X, \hat{m}(s)\big)$. We upper-bound $\underline{h}_s(X) - \underline{h}_s(Y)$ by $h_s\big(X, \hat{m}(s)\big) - h_s\big(Y, \hat{m}(s)\big)$, and use the α-contraction property of h to derive

$$\begin{aligned}\underline{h}_s(X) - \underline{h}_s(Y) &\leqslant |h_s\big(X, \hat{m}(s)\big) - h_s\big(Y, \hat{m}(s)\big)| \\ &\leqslant \alpha\|X - Y\|_\infty.\end{aligned}$$

Since X and Y are arbitrarily ordered, we conclude that $\|\underline{h}(X) - \underline{h}(Y)\|_\infty \leqslant \alpha\|X - Y\|_\infty$. The proof for $\overline{h}$ follows a similar reasoning and takes $\hat{m}(s) = \operatorname{argmax}_{m\in\mathcal{M}} h_s(X, m)$. The existence of $\underline{X}(\overline{X})$ follows from applying Banach's fixed point theorem. □

Lemma 4 (*Order Preservation*)**.** *The bound operators* $\underline{h}$ *and* $\overline{h}$ (23) *are order-preserving on* $\mathbb{R}^S$ (*Definition* 2), *i.e.,*

$$\forall\, U, V \in \mathbb{R}^S,\ U \leqslant V \Rightarrow \underline{h}(U) \leqslant \underline{h}(V),\ \overline{h}(U) \leqslant \overline{h}(V).$$

Proof. The lemma statement follows directly from the fact that order preservation is conserved through composition with inf and

sup. If $h(U,m) \leqslant h(V,m)$, then $\inf_{m\in\mathcal{M}} h(U,m) \leqslant \inf_{m\in\mathcal{M}} h(V,m)$. A similar argument follows for $\overline{h}(\cdot) = \sup_{m\in\mathcal{M}} h(\cdot,m)$. □

We show that the fixed points $\underline{X}$ and $\overline{X}$ bound the fixed point set $\mathcal{V}^\star$ of the set-based value operator H (14).

Theorem 2 (*Bounding Fixed Point Sets*)**.** *If h (6) is a value operator on $\mathbb{R}^S \times \mathcal{M}$ and $\mathcal{M}$ is compact,*

$$\underline{X} \leqslant V \leqslant \overline{X},\ \forall\, V \in \mathcal{V}^\star, \tag{25}$$

where $\underline{X}$ and $\overline{X}$ (24) are the fixed points of the bound operators $\underline{h}$ and $\overline{h}$ (23). Here, $\mathcal{V}^\star$ is the fixed point set of the set-based value operator H (14) induced by h (6) on $\mathbb{R}^S \times \mathcal{M}$.

Proof. For $\mathcal{V}^0 = \{\underline{X}, \overline{X}\}$ and $\mathcal{V}^{k+1} = H(\mathcal{V}^k)$ (20), we first show

$$\underline{X} \leqslant V \leqslant \overline{X},\ \forall\, V \in \mathcal{V}^k, \tag{26}$$

via induction. Suppose that (26) is satisfied for $\mathcal{V}^k$. The order preserving property of $h(\cdot,m)$ implies that $h(\underline{X},m) \leqslant h(V,m) \leqslant h(\overline{X},m)$ holds for all $(V,m) \in \mathcal{V}^k \times \mathcal{M}$. We take the infimum and supremum over $h(\underline{X},m)$ and $h(\overline{X},m)$, respectively, to show that for all $(V,m) \in \mathcal{V}^k \times \mathcal{M}$ and $s \in [S]$,

$$\inf_{m'\in\mathcal{M}} h_s(\underline{X},m') \leqslant h_s(V,m) \leqslant \sup_{m'\in\mathcal{M}} h_s(\overline{X},m').$$

Since $\underline{X}$ and $\overline{X}$ are the fixed points of $\inf_{m'\in\mathcal{M}} h_s(\cdot,m')$ and $\sup_{m'\in\mathcal{M}} h_s(\cdot,m')$ for all $s \in [S]$, respectively, we conclude that (26) holds for $\mathcal{V}^{k+1}$.

Next, we show that $\underline{X}$ and $\overline{X}$ bound the fixed point set $\mathcal{V}^\star$ for the h-induced operator H (14). From Lemma 5, we know that for all $V \in \mathcal{V}^\star$, there exists a strictly increasing sequence $\phi : \mathbb{N} \mapsto \mathbb{N}$ and corresponding value vectors $\{W^{\phi(n)}\}$ such that $\lim_{n\to\infty} W^{\phi(n)} = V$ and $W^{\phi(n)} \in \mathcal{V}^{\phi(n)}$ for the sequence of value vector sets generated from $\mathcal{V}^0 = \{\underline{X}, \overline{X}\}$. Since $\underline{X} \leqslant W^{\phi(n)} \leqslant \overline{X}$ holds for all n, we conclude (25) holds. □

5. Revisiting min–max MDP

The extremum of the set-based Bellman operator's fixed point set is equivalent to the min–max value vector under the rectangularity condition, though it exists under fewer restrictions. We formally prove this relationship by re-examining robust MDPs using a set-theoretic approach. Recall the optimistic value vector $W^o \in \mathbb{R}^S$ and robust value vector $W^r \in \mathbb{R}^S$ of a discounted MDP $([S],[A],C,P,\gamma)$ from Iyengar (2005) and Nilim and El Ghaoui (2005) as the fixed points of the following operators.

$$W_s^o = \min_{\pi_s\in\Delta_A} \min_{(C,P)\in\mathcal{M}} g_s^\pi(W^o,C,P),\ \forall s \in [S], \tag{27}$$

$$W_s^r = \min_{\pi_s\in\Delta_A} \max_{(C,P)\in\mathcal{M}} g_s^\pi(W^r,C,P), \forall s \in [S]. \tag{28}$$

The optimistic policy π^o and robust policy π^r are the optimal policies corresponding to (27) and (28), respectively.

$$\pi_s^o \in \operatorname*{argmin}_{\pi_s\in\Delta_A} \min_{(C,P)\in\mathcal{M}} g_s^\pi(W^o,C,P), \forall s \in [S] \tag{29}$$

$$\pi_s^r \in \operatorname*{argmin}_{\pi_s\in\Delta_A} \max_{(C,P)\in\mathcal{M}} g_s^\pi(W^r,C,P), \forall s \in [S] \tag{30}$$

For clarity, we denote the policy evaluation operator (7) under π^o as g^o and the policy evaluation operator (7) under π^r as g^r.

When $\mathcal{M}$ is (s,a)-rectangular (35), the set of policies satisfying (29) and (30) is non-empty and includes deterministic policies (Iyengar, 2005, Thm 3.1). When $\mathcal{M}$ is s-rectangular and convex, the set of policies satisfying (30) is non-empty but may be mixed (Wiesemann et al., 2013, Thm 4). When $\mathcal{M}$ is convex, we show that policies (29) and (30) exist.

Proposition 2. *If the MDP parameter set $\mathcal{M}$ is compact and convex, then*

(1) W^o (27) and W^r (28) exist and satisfy $\overline{f}(W^r) = W^r$, $\underline{f}(W^o) = W^o$, where $\overline{f}$ and $\underline{f}$ (23) are the bound operators of the Bellman operator (9).
(2) π^o (29) and π^r (30) exist.

Proof. Recall the Bellman operator f (9). When $\mathcal{M} \times \Delta_A$ is compact, the formulation of the fixed point of $\underline{f}$ (23) is equivalently given by

$$\underline{f}_s(\underline{X}) = \min_{(C,P)\in\mathcal{M}} \min_{\pi_s\in\Delta_A} g_s^\pi(\underline{X},C,P),\ \forall s \in [S]. \tag{31}$$

We note that (31) is identical to the formulation of W^o (27). Therefore, $W^o = \underline{X}$ is the fixed point of $\underline{f}$. When $\mathcal{M}$ is compact, W^o exists due to Lemma 3. From (29), π_s^o is the optimal argument of $g_s^\pi(W^o,C,P)$, a continuous function in π_s, C, P minimized over compact sets $\Delta_A \times \mathcal{M}$ for all $s \in [S]$. Therefore π_s^o exists. Since $\pi^o = (\pi_1^o, \ldots, \pi_S^o)$, the optimal $\pi^o \in \Delta_A^S$ exists.

For the robust scenario: when $\mathcal{M}$ is compact, the fixed point of $\overline{f}$ (23), $\overline{X}$, exists from Lemma 3 and is given by

$$\overline{X}_s = \max_{(C,P)\in\mathcal{M}} \min_{\pi_s\in\Delta_A} g_s^\pi(\overline{X},C,P),\ \forall s \in [S]. \tag{32}$$

The function $g_s^\pi(\overline{X},C,P)$ is concave in (C,P) and convex in π. If $\mathcal{M}$ is convex, then we apply the minimax theorem (Neumann, 1928) to switch the order of min and max in (32) to derive

$$\overline{X}_s = \min_{\pi_s\in\Delta_A} \max_{(C,P)\in\mathcal{M}} g_s^\pi(\overline{X},C,P),\ \forall s \in [S]. \tag{33}$$

Eq. (33) is identical to (28), therefore $W^r = \overline{X}$ and exists by Lemma 3. In (33), $\max_{(C,P)\in\mathcal{M}} g_s^\pi(\overline{X},C,P)$ is piece-wise linear in π_s and Δ_A is compact for all $s \in [S]$, thus $\operatorname{argmin}_{\pi_s\in\Delta_A} \max_{(C,P)\in\mathcal{M}} g_s^\pi(\overline{X},C,P)$ is non-empty. Finally, since $\pi^r = (\pi_1^r, \ldots, \pi_S^r)$, π^r exists. □

Remark 6. Since $\max_{(C,P)\in\mathcal{M}} g_s^\pi(\overline{X},C,P)$ is piecewise linear in π_s, the optimal π_s^r is mixed in general. This is consistent with the results in Wiesemann et al. (2013).

Proposition 2 generalizes the results from Wiesemann et al. (2013) to show that (28) exists when $\mathcal{M}$ is compact and convex instead of s-rectangular and convex. In particular, if we construct $\hat{\mathcal{M}} = \prod_{s\in[S]} \operatorname{proj}_s(\mathcal{M})$, where $\operatorname{proj}_s(\mathcal{M})$ is the projection of the elements of $\mathcal{M}$ onto the s coordinate, then the policies π_o and π_r can be computed using the robust policy iteration from Wiesemann et al. (2013). Diverging from the min–max approach, W^o and W^r always exist and do not require the rectangularity conditions (Iyengar, 2005).

5.1. Containment and rectangularity

Since the fixed point set $\mathcal{V}^\star$ is a subset of a multi-dimensional vector space $\mathbb{R}^S$, it is possible that $\mathcal{V}^\star$ does not contain its own extremal elements. We illustrate this in Fig. 5. We also show that interestingly, MDP rectangularity is a sufficient condition for the fixed point set to contain its own extrema.

Fig. 5 implies that the structure of $\mathcal{M}$ may dictate whether the fixed point set contains its own extrema points. We formally prove this in the following section.

Assumption 1 (*Containment Condition*)**.** The parameter uncertainty set $\mathcal{M}$ satisfies the containment condition with respect to h if $\mathcal{M}$ is compact and for all $V \in \mathbb{R}^S$,

$$\bigcap_{s\in[S]} \operatorname*{argmin}_{m\in\mathcal{M}} h_s(V,m) \neq \varnothing,\ \bigcap_{s\in[S]} \operatorname*{argmax}_{m\in\mathcal{M}} h_s(V,m) \neq \varnothing. \tag{34}$$

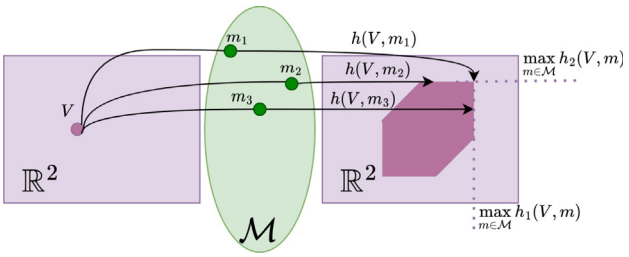


Fig. 5. The $\operatorname{argmax}_{m\in\mathcal{M}} h_s(V, m)$ for a value operator h given $\mathcal{M}$ when $S = 2$. Here, $\operatorname{argmax}_{m\in\mathcal{M}} h_1(V, m) = \{m_1, m_3\}$, $\operatorname{argmax}_{m\in\mathcal{M}} h_2(V, m) = \{m_1, m_2\}$. Therefore, $\mathcal{M}$ satisfies Assumption 1.

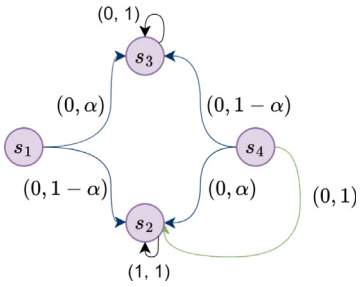


Fig. 6. MDP with parameter coupling in transition probability across different states.

Assumption 1 states that (1) there exists a parameter m such that $h(\cdot, m)$'s fixed point is the extremum of $\mathcal{V}^\star$ and (2) $m \in \mathcal{M}$. Intuitively, if there exists $m \in \mathcal{M}$ such that $V = h(V, m)$, then from Corollary 1, $V \in \mathcal{V}^\star$.

Remark 7. Assumption 1 is an h-dependent condition imposed on the structure of $\mathcal{M}$, and is independent of $\mathcal{M}$'s convexity and connectivity.

With respect to the Bellman operator f (9) and the policy evaluation operators g^π (7), the following conditions in min–max MDP are sufficient to satisfy Assumption 1.

Definition 11 (*(s, a)-Rectangular Sets Iyengar, 2005; Nilim & El Ghaoui, 2005*)**.** The uncertainty set $\mathcal{M} \subset \mathbb{R}^{S\times A} \times \Delta_S^{SA}$ is (s, a)-rectangular if

$$\mathcal{M} = \bigtimes_{(s,a)\in[S]\times[A]} \mathcal{M}_{sa},\ \mathcal{M}_{sa} \subset \mathbb{R} \times \Delta_S,\ \forall (s, a) \in [S] \times [A]. \tag{35}$$

Intuitively, (s, a)-rectangularity implies that the MDP parameter uncertainty is *decoupled* between each state–action.

Definition 12 (*s-Rectangular Sets*)**.** The uncertainty set $\mathcal{M} \subset \mathbb{R}^{S\times A} \times \Delta_S^{SA}$ is s-rectangular if

$$\mathcal{M} = \bigtimes_{s\in[S]} \mathcal{M}_s,\ \mathcal{M}_s \subset \mathbb{R}^A \times \Delta_S^A,\ \forall s \in [S]. \tag{36}$$

Remark 8. Definition 12 applies to ambiguity sets that arise in distributionally robust MDPs (Xu & Mannor, 2010; Yang, 2017; Yu & Xu, 2015).

Example 2 (*Wind Uncertainty*)**.** Consider the navigation problem presented in Example 1. If the wind pattern strictly switches between N wind patterns, then the transition uncertainty at state $s \in [S]$ is given by $\mathcal{P}_s = \{P_s^1, \ldots, P_s^N\}$. If the wind pattern is a mixture of N discrete wind trends, the transition uncertainty at state $s \in [S]$ is $\mathcal{P}_s = \{\sum_i \alpha_i P_s^i \mid \alpha \in \Delta_N\}$. Both wind patterns lead to s-rectangular uncertainty, given by $\mathcal{P} = \bigtimes_{s\in[S]} \mathcal{P}_s$.

We show that the rectangularity conditions indeed are sufficient for satisfying Assumption 1 with respect to f (9) and g^π (7).

Proposition 3. *If $\mathcal{M}$ is compact and s-rectangular (Definition 12), $\mathcal{M}$ satisfies Assumption 1 with respect to f (9) and g^π (7) for all $\pi \in \Delta_A^S$.*

Proof. We first show that $\mathcal{M}$ satisfies Assumption 1 with respect to the Bellman operator. Given $s \in [S]$, $f_s(V, C, P)$ only depends on the s component of C and P. From Lemma 1, f_s is continuous in (c_s, P_s). Let $(c_s^\star, P_s^\star)$ be the solution to $\operatorname{argmin}_{(c_s,P_s)\in\mathcal{M}_s} f_s(V, C, P)$ for all $\forall\, s \in [S]$. If $\mathcal{M}_s$ is compact, $(c_s^\star, P_s^\star) \in \mathcal{M}_s$. We can construct $C^\star = [c_1^\star, \ldots, c_S^\star]$ and $P^\star = [P_1^\star, \ldots, P_S^\star]$. If $\mathcal{M}$ is s-rectangular, then $(C^\star, P^\star) \in \mathcal{M}$ and $(C^\star, P^\star) \in \operatorname{argmin}_{(C,P)\in\mathcal{M}} f_s(V, C, P)$ for all $s \in [S]$. We conclude that $\mathcal{M}$ satisfies Assumption 1.

Given $\pi \in \Delta_A^S$ and $s \in [S]$, g_s^π only depends on c_s and P_s as well. We can similarly show that there exists an optimal parameter $(C^\star, P^\star) \in \operatorname{argmin}_{(C,P)\in\mathcal{M}} g_s^\pi(V, C, P)$ for all $s \in [S]$ such that $(C^\star, P^\star) \in \mathcal{M}$. □

Beyond s-rectangularity, there are sets that satisfy Assumption 1 with respect to specific value operators.

Example 3 (*Beyond Rectangularity*)**.** Consider a four state MDP with α-parametrized transition uncertainty $\mathcal{M}$ in Fig. 6, where states are the nodes and actions are the multi-headed arrows. Each head has an associated tuple $(C_{sa}, P_{sa,s'})$ denoting its cost and transition probability. The states s_2 and s_3 have values $V_2 = \frac{1}{1-\gamma}$ and $V_3 = 0$ for both f and g^π for all $\pi \in \Delta_A^S$.

The states s_1 and s_4 have transition uncertainty jointly parametrized by $\alpha \in [0, 1]$, therefore violating s-rectangularity (Definition 12). The optimal cost-to-go values V_1 and V_4 occur at different α's, therefore violating Assumption 1 with respect to f. However, suppose that at s_4, we only choose the action with cost-transition $(0, 1)$ to s_2 in Fig. 6. Then V_4 is independent of α. The minimum and maximum V_1 occur at $\alpha = 1$ and $\alpha = 0$, respectively. Therefore, $\mathcal{M}$ satisfies Assumption 1 with respect to operator g^π for all $\pi = [\pi_{s_1}, \ldots, \pi_{s_4}]$ where $\pi_{s_4} = [1, 0]$.

When Assumption 1 is satisfied, the fixed point of H (14) contains its own supremum and infimum value vectors.

Theorem 3. *If h (6) on $\mathbb{R}^S \times \mathcal{M}$ satisfies Assumption 1, then there exists $\underline{m}, \overline{m} \in \mathcal{M}$ such that $\overline{h}$ and $\underline{h}$ (23) and their fixed points $\underline{X}$ and $\overline{X}$ (24) satisfies*

$$\underline{h}(\underline{X}) = h(\underline{X}, \underline{m}) = \underline{X},\ \ \overline{h}(\overline{X}) = h(\overline{X}, \overline{m}) = \overline{X}. \tag{37}$$

Additionally, $\underline{X}$ and $\overline{X}$ are the least and the greatest elements of H's fixed point set $\mathcal{V}^\star$, $\underline{V}^\star, \overline{V}^\star$ (22) respectively, and both belong to $\mathcal{V}^\star$ (17).

$$\underline{X} = \underline{V}^\star,\ \ \overline{X} = \overline{V}^\star,\ \ \underline{X}, \overline{X} \in \mathcal{V}^\star.$$

Proof. From Theorem 2, $\underline{X}$ and $\overline{X}$ are the lower and upper bounds on the fixed point set $\mathcal{V}^\star$. We show that these are the infimum and supremum elements of $\mathcal{V}^\star$ by showing that they are also elements of $\mathcal{V}^\star$. From Assumption 1, there exists $\underline{m}, \overline{m} \in \mathcal{M}$ such that $h_s(\underline{X}, \underline{m}) = \min_{m\in\mathcal{M}} h_s(\underline{X}, \underline{m})$ and $h_s(\overline{X}, \overline{m}) = \min_{m\in\mathcal{M}} h_s(\overline{X}, \overline{m})$ for all $s \in [S]$. Since $\underline{X}$ and $\overline{X}$ are fixed points of $h(\cdot, \underline{m})$ and $h(\cdot, \overline{m})$, we apply Corollary 1 to conclude that $\underline{X}, \overline{X} \in \mathcal{V}^\star$. □

5.2. Performance of non-stationary Bellman update

Consider the wind navigation problem in Example 1, one question posed is whether it is better to use the robust policy, optimistic policy, or the next step optimal policy. In this section, we use the set-theoretic tools to compare the performance of these approaches. First, we introduce some notations: let $G^o = G^{\pi_o}$ (7), the fixed point of G^o be $\mathcal{V}^o$, $G^r = G^{\pi^r}$, and the fixed point of G^r be $\mathcal{V}^r$.

$$\mathcal{V}^o = \{g^o(V, C, P) \mid (C, P) \in \mathcal{M}, V \in \mathcal{V}^o\}, \tag{38}$$

$$\mathcal{V}^r = \{g^r(V, C, P) \mid (C, P) \in \mathcal{M}, V \in \mathcal{V}^r\}. \tag{39}$$

Additionally, the supremum value vectors of $\mathcal{V}^o$ and $\mathcal{V}^r$ are $\overline{V}^o$ and $\overline{V}^r$ respectively and the infimum value vectors are $\underline{V}^o$ and $\underline{V}^r$, respectively.

$$\underline{V}^r_s = \min_{V \in \mathcal{V}^r} V_s, \ \overline{V}^r_s = \max_{V \in \mathcal{V}_r} V_s, \ \forall s \in [S]. \tag{40}$$

$$\underline{V}^o_s = \min_{V \in \mathcal{V}^o} V_s, \ \overline{V}^o_s = \max_{V \in \mathcal{V}^o} V_s, \ \forall s \in [S]. \tag{41}$$

We compare these with the fixed point set of the Bellman operator, $\mathcal{V}^B = \{\min_\pi g^\pi(V, C, P) \mid (C, P) \in \mathcal{M}, V \in \mathcal{V}^B\}$ (17), denoted by $\overline{V}^B$ and $\underline{V}^B$ as

$$\underline{V}^B_s = \min_{V \in \mathcal{V}^B} V_s, \ \overline{V}^B_s = \max_{V \in \mathcal{V}^B} V_s, \ \forall s \in [S]. \tag{42}$$

Theorem 4. *If f, g^o, g^r satisfy Assumption 1 on $\mathbb{R}^S \times \mathcal{M}$, then the bounding value vectors (42) (41) (40) of the corresponding fixed point sets $\mathcal{V}^B$, $\mathcal{V}^o$ (38) and $\mathcal{V}^r$ (39) are ordered as*

$$\underline{V}^B = \underline{V}^o \leqslant \underline{V}^r, \ \overline{V}^B = \overline{V}^r \leqslant \overline{V}^o. \tag{43}$$

Proof. Since $\underline{V}^o$ is the infimum element for the fixed point set $\mathcal{V}^o$ (41), we can apply Theorem 3 to derive

$$\underline{V}^o = \min_{(C,P)\in\mathcal{M}} g^o(\underline{V}^o, C, P). \tag{44}$$

By definition of π^o (29), $\min_{(C,P)\in\mathcal{M}} g^o(\underline{V}^o, C, P) = \min_{(C,P)\in\mathcal{M}} \min_{\pi\in\Delta_A^S} g^\pi(\underline{V}^o, C, P)$. As the two minima commute,

$$\min_{(C,P)\in\mathcal{M}} g^o(\underline{V}^o, C, P) = \min_{(C,P)\in\mathcal{M}} \min_{\pi\in\Delta_A^S} g^\pi(\underline{V}^o, C, P). \tag{45}$$

Combining (44) and (45), $\underline{V}^o$ is exactly the unique fixed point of $\min_{(C,P)\in\mathcal{M}} \min_{\pi\in\Delta_A^S} g^\pi(\cdot, C, P)$. However, by applying Theorem 3 to f on $\mathbb{R}^S \times \mathcal{M}$, $\underline{V}^B$ is also the unique fixed point of $\min_{(C,P)\in\mathcal{M}} \min_{\pi\in\Delta_A^S} g^\pi(\cdot, C, P)$. Therefore $\underline{V}^o = \underline{V}^B$.

From (40), $\underline{V}^r = \min_{(C,P)\in\mathcal{M}} g^r(\underline{V}^r, C, P)$, we can minimize over the policy space to lower bound $\underline{V}^r$ as

$$\underline{V}^r \geqslant \min_{\pi\in\Delta_A^S} \min_{(C,P)\in\mathcal{M}} g^r(\underline{V}^r, C, P). \tag{46}$$

Since the right hand side of (46) is equivalent to $\underline{f}(\underline{V}^r)$, (46) is equivalent to $\underline{V}^r \geqslant \underline{f}(\underline{V}^r)$. From Lemma 4, $\underline{f}$ is order-preserving in V, we conclude that $\underline{V}^o = \underline{V}^\star \leqslant \underline{V}^r$.

From Theorem 3, $\overline{V}^r$ is the fixed point of $\overline{g}^r$, such that

$$\overline{V}^r = \max_{(C,P)\in\mathcal{M}} g^r(\overline{V}^r, C, P). \tag{47}$$

We apply $\min_\pi$ to both sides of (47) and use the definition of π^r to derive that $\overline{V}^r$ is the fixed point of $\min_{\pi\in\Delta_S^A} \max_{(C,P)\in\mathcal{M}} g^\pi(V^r, C, P)$. From Assumption 1, there exists $(\overline{C}, \overline{P}) \in \mathcal{M}$ that maximizes $g^\pi(\overline{V}, C, P)$, so $\overline{V}^r$ equivalently satisfies

$$\overline{V}^r = \min_{\pi\in\Delta_A^S} g^\pi(\overline{V}^r, \overline{C}, \overline{P}).$$

From Corollary 1, $\overline{V}^r \in \mathcal{V}^B$ and therefore $\overline{V}^r \leqslant \overline{V}^B$. Next we show $\overline{V}^B \leqslant \overline{V}^r$. From Theorem 3, $\overline{V}^B$ is the fixed point of $\overline{f}$, such that

$$\overline{V}^B = \max_{(C,P)\in\mathcal{M}} \min_\pi g^\pi(\overline{V}^B, C, P),$$

From the min–max inequality,

$$\overline{V}^B \leqslant \min_{\pi\in\Delta_A^S} \max_{(C,P)\in\mathcal{M}} g^\pi(\overline{V}^B, C, P).$$

Since $\pi^r \in \Delta_A^S$,

$$\overline{V}^B \leqslant \max_{(C,P)\in\mathcal{M}} g^r(\overline{V}^B, C, P). \tag{48}$$

The right-hand side of (48) is $\overline{g}^r(\overline{V}^B)$ (23), such that (48) is equivalent to $\overline{V}^B \leqslant \overline{g}^r(\overline{V}^B)$. Consider the sequence $V^{k+1} = \overline{g}^r(V^k)$ where $V^1 = \overline{V}^B$. Since $\overline{g}^r$ is a contraction, $\lim_{k\to\infty} V^k = V^r$, the fixed point of $\overline{g}^r$. From Lemma 4, $\overline{g}^r$ is order preserving. Therefore $\overline{V}^B = V^1 \leqslant V^r$.

Finally, Theorem 3 implies that $\overline{V}^o$ is the fixed point of $\overline{g}^o$: $\overline{V}^o = \max_{(C,P)\in\mathcal{M}} g^o(\overline{V}^o, C, P)$. By construction, $\overline{V}^o \geqslant \min_{\pi\in\Delta_A^S} \max_{(C,P)\in\mathcal{M}} g^\pi(\overline{V}^o, C, P)$. From the min–max inequality,

$$\min_{\pi\in\Delta_A^S} \max_{(C,P)\in\mathcal{M}} g^\pi(\overline{V}^o, C, P) \geqslant \max_{(C,P)\in\mathcal{M}} \min_{\pi\in\Delta_A^S} g^\pi(\overline{V}^o, C, P),$$

such that the right hand side of the inequality is equivalent to $\overline{f}(\overline{V}^o)$. Following the monotonicity properties of the Bellman operator f (Puterman, 2014, Thm. 6.2.2), we conclude that $\overline{V}^o \geqslant \overline{V}^B$. □

Remark 9. Our set-theoretic approach shows that in addition to having the best worst-case performance among $\{\mathcal{V}^o, \mathcal{V}^B, \mathcal{V}^r\}$, $\mathcal{V}^r$ also has the smallest variation in performance for the same uncertainty set $\mathcal{M}$.

Recall Example 1, iteratively updating policy using the next step cost and transition dynamics will result in a value vector trajectory that asymptotically converges to $\mathcal{V}^B$. In that case, Theorem 4 implies this policy update scheme has comparable performance to the optimistic policy in non-adversarial wind fields but will perform no worse than the robust policy in adversarial winds.

Finally, we generalize the *s*-rectangularity condition by showing that the optimistic and robust policies exist when the MDP parameter set $\mathcal{M}$ satisfies Assumption 1.

Corollary 3 (*Robust MDP under Assumption 1*)**.** *If $\mathcal{M}$ is compact, convex, and f, g^o, g^r satisfy Assumption 1 on $\mathbb{R}^S \times \mathcal{M}$, then W^o (27) and W^r (28) are the infimum and supremum value vectors of $\mathcal{V}^o$ and $\mathcal{V}^r$,*

$$W^o_s = \inf_{V\in\mathcal{V}^o} V_s, W^r_s = \sup_{V\in\mathcal{V}^r} V_s, \ \forall s \in [S], \tag{49}$$

where $\mathcal{V}^o$ (38) and $\mathcal{V}^r$ (39) are the fixed point sets of G^o and G^r, respectively (see Fig. 7).

Proof. When f satisfies Assumption 1 on $\mathbb{R}^S \times \mathcal{M}$, Theorem 3 shows that $\underline{V}^B = W^o$, $\overline{V}^B = W^r$. If g^o, and g^r also satisfies Assumption 1 on $\mathbb{R}^S \times \mathcal{M}$, then we apply Theorem 4 to derive $W^o = \underline{V}^o$ and $W^r = \overline{V}^r$. This proves the corollary statement. □

Remark 10. When Assumption 1 is not satisfied, W^o and W^r still bound $\underline{V}^o$ and $\overline{V}^r$.

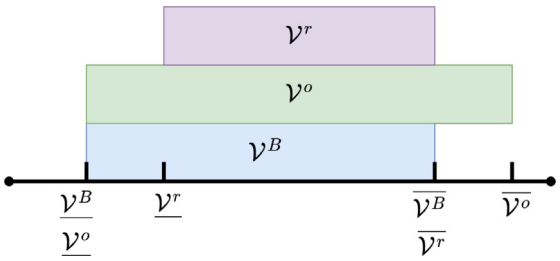


Fig. 7. Visualization of Theorem 4. The purple, green, blue regions indicate the ranges of $\mathcal{V}^r$, $\mathcal{V}^o$, and $\mathcal{V}^B$, respectively.

6. Value iteration for fixed point set computation

In the previous sections, we proved the existence of a fixed point set for value operators with compact parameter uncertainty sets and re-interpreted robust control. Next, we derive an iterative algorithm for computing the bounds of the fixed point set $\mathcal{V}$ given a value operator h and parameter uncertainty set $\mathcal{M}$.

Algorithm Sketch. Based on the set-based value iteration (20), we iteratively find the one-step bounds of $H(\mathcal{V}^k)$ to converge the bounds of the fixed point set.

For any compact set $\mathcal{V} \in \mathcal{K}(\mathbb{R}^S)$, the one step bounds of $H(\mathcal{V})$ are the result of applying the one-step bound operators $\underline{h}$ and $\overline{h}$ (23) to the extremal points of $\mathcal{V}$.

Theorem 5 (*One Step H Bounds*)**.** *Consider the set operator H (14) and its bound operators $\underline{h}$ and $\overline{h}$ (23) induced by h on $\mathbb{R}^S \times \mathcal{M}$ (6). For a compact set $\mathcal{V} \subset \mathbb{R}^S$, $H(\mathcal{V})$ is bounded by $\underline{h}(\underline{V})$ and $\overline{h}(\overline{V})$ (23) as*

$$\underline{h}(\underline{V}) \leqslant V \leqslant \overline{h}(\overline{V}), \quad \forall\, V \in H(\mathcal{V}). \tag{50}$$

where $\underline{V}$ and $\overline{V}$ (22) are the extremal elements of $\mathcal{V}$. If h satisfies Assumption 1 on $\mathbb{R}^S \times \mathcal{M}$ and $\underline{V}, \overline{V} \in \mathcal{V}$, then $\underline{h}(\underline{V})$ and $\overline{h}(\overline{V})$ are the supremum and infimum elements of $H(\mathcal{V})$, respectively—for all $s \in [S]$, $\underline{h}_s(\underline{V})$ and $\overline{h}_s(\overline{V})$ satisfy

$$\underline{h}_s(\underline{V}) = \inf_{(V,m)\in\mathcal{V}\times\mathcal{M}} h_s(V,m), \ \overline{h}_s(\overline{V}) = \sup_{(V,m)\in\mathcal{V}\times\mathcal{M}} h_s(V,m). \tag{51}$$

Proof. For all $s \in [S]$, $h_s(V,m) \leqslant \overline{h}_s(V)$ for all $m \in \mathcal{M}$. If h is $K(V)$-Lipschitz and α-contractions in $\mathcal{M}$, then $\overline{h}$ is order-preserving (Lemma 4) such that $\overline{h}_s(V) \leqslant \overline{h}_s(\overline{V})$ for all $V \in \mathcal{V}$. We conclude that

$$h(V,m) \leqslant \overline{h}(\overline{V}), \ \forall (V,m) \in \mathcal{V} \times \mathcal{M}. \tag{52}$$

Since $\overline{h}$ is an upper bound, and $\sup$ is the least upper bound, it holds that $\sup_{V,m} h_s(V,m) \leqslant \overline{h}(\overline{V})$. We use the definition of $H(\mathcal{V})$ (14) to conclude that $V \leqslant \overline{h}(\overline{V})$ for all $V \in H(\mathcal{V})$. The inequality $\underline{h}(\underline{V}) \leqslant V \ \forall V \in H(\mathcal{V})$ can be similarly proved.

If h satisfies Assumption 1 on $\mathbb{R}^S \times \mathcal{M}$ and $\underline{V}, \overline{V} \in \mathcal{V}$, Assumption 1 states that there exists $\underline{m} \in \mathcal{M}$ such that $h(\underline{V}, \underline{m}) = \underline{h}(\underline{V})$. Therefore, $\underline{h}(\underline{V}) \in H(\mathcal{V})$. Since $\underline{h}(\underline{V})$ also lower bounds all the elements of $H(\mathcal{V})$, it is the infimum element of $H(\mathcal{V})$. The fact that the greatest element of $H(\mathcal{V})$ is $\overline{h}(\overline{V})$ can be similarly proved. □

Based on Theorem 5, we propose the following bound approximation algorithm of the fixed point set $\mathcal{V}^\star$ (17) for a set-valued operator H (6).

6.1. Computing one-step optimal parameters

Algorithm 1 is stated for a general MDP parameter set $\mathcal{M}$ and does not specify how to compute lines 4 and 5. Here are some solution methods for different types of $\mathcal{M}$.

Algorithm 1 Bounding the fixed point set $\mathcal{V}$

Input: $\mathcal{C}$, $\mathcal{P}$, V^0, ϵ.
Output: $\underline{V}$, $\overline{V}$
1: $\underline{V}^0 := \overline{V}^0 := V^0$
2: $e^0 = \frac{1-\gamma}{\gamma}\epsilon$
3: **while** $\frac{\gamma}{1-\gamma} e^k \geqslant \epsilon$ **do**
4: $\quad \underline{V}_s^{k+1} = \min_{m\in\mathcal{M}} h_s(\underline{V}^k, m), \quad \forall s \in [S]$
5: $\quad \overline{V}_s^{k+1} = \max_{m\in\mathcal{M}} h_s(\overline{V}^k, m), \quad \forall s \in [S]$
6: $\quad e^{k+1} = \max\left\{\|\underline{V}^{k+1} - \underline{V}^k\|, \|\overline{V}^{k+1} - \overline{V}^k\|\right\}$
7: $\quad k = k+1$
8: **end while**

(1) **Finite** $\mathcal{M}$. If $\mathcal{M} = \{m_1, \ldots, m_N\}$ has a finite number of elements, we can directly compute line 4 as

$$\underline{V}^{k+1} = \min\left\{h_s(\underline{V}^k, m_i) \mid i = \{1, \ldots, N\}\right\}. \tag{53}$$

For line 5, we replace min with max in (53).

(2) **Convex** $\mathcal{M}$. When $\mathcal{M}$ is a convex set, the computation depends on h. If $h = g^\pi$ is the policy operator, lines 4 and 5 can be solved as convex optimization problems. If h is the Bellman operator f, lines 4 and 5 take on min–max formulation and is NP-hard to solve in the general form Wiesemann et al. (2013). When $\mathcal{M}$ can be characterized by an ellipsoidal set of parameters, the solutions to lines 4 and 5 is given in Wiesemann et al. (2013).

6.2. Algorithm convergence rate

When lines 4 and 5 are solvable, Algorithm 1 asymptotically converges to approximations of the bounding elements of $\mathcal{V}^\star$. If $\mathcal{M}$ satisfies Assumption 1 with respect to h, Algorithm 1 derives the exact bounds of $\mathcal{V}^\star$. Algorithm 1 has similar rates of convergence in Hausdorff distance as standard value iteration using h on $\mathbb{R}^S$.

Theorem 6. *Consider the value operator h, compact uncertainty set $\mathcal{M}$, and the fixed point set $\mathcal{V}^\star$ of the set-based operator H (14) induced by h on $\mathbb{R}^S \times \mathcal{M}$. If $\mathcal{M}$ satisfies Assumption 1 with respect to h, then at each iteration k,*

$$\begin{aligned} \|\underline{V}^{k+1} - \underline{V}^\star\|_\infty &\leqslant \alpha \|\underline{V}^k - \underline{V}^\star\|_\infty, \\ \|\overline{V}^{k+1} - \overline{V}\|_\infty &\leqslant \alpha \|\overline{V}^k - \overline{V}^\star\|_\infty, \end{aligned} \tag{54}$$

where all norms are infinity norms, and $\underline{V}^\star, \overline{V}^\star$ are the infimum and supremum bounds of $\mathcal{V}$, respectively. At Algorithm 1's termination, $\underline{V}^k, \overline{V}^k$ satisfies

$$\max\left\{\|\underline{V}^k - \underline{V}^\star\|_\infty, \|\overline{V}^k - \overline{V}^\star\|_\infty\right\} < \epsilon. \tag{55}$$

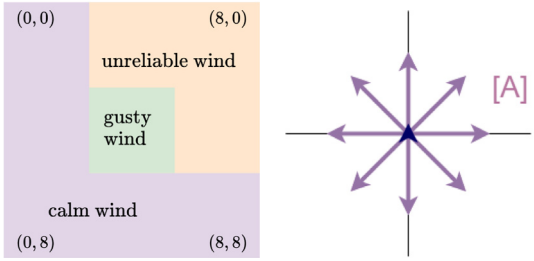
Proof. From Algorithm 1, $\overline{V}^{k+1} = \overline{h}(\overline{V}^k)$. From Lemma 3, $\overline{h}$ is an α-contraction. We obtain

$$\|\overline{V}^{k+1} - \overline{V}^\star\|_\infty \leqslant \alpha \|\overline{V}^k - \overline{V}^\star\|_\infty$$

and note that (54) holds by induction. Next, we apply triangle inequality to $\|\overline{V}^k - \overline{V}^\star\|_\infty$ as

$$\|\overline{V}^k - \overline{V}^\star\|_\infty \leqslant \|\overline{V}^k - \overline{V}^{k+1}\|_\infty + \|\overline{V}^{k+1} - \overline{V}^\star\|_\infty. \tag{56}$$

We can then use $\|\overline{V}^{k+1} - \overline{V}^\star\|_\infty \leqslant \alpha \|\overline{V}^k - \overline{V}^\star\|_\infty$ to bound (56) as $\|\overline{V}^k - \overline{V}^\star\|_\infty \leqslant \frac{1}{1-\alpha}\|\overline{V}^k - \overline{V}^{k+1}\|_\infty$. A similar argument can show that $\|\underline{V}^k - \underline{V}^\star\|_\infty \leqslant \frac{1}{1-\alpha}\|\underline{V}^k - \underline{V}^{k+1}\|_\infty$. When Algorithm 1's while



(a) Wind field discretized into 81 states.

(b) Nine actions representing different thrusts per state.

Fig. 8. Wind field MDP state space and action space.

condition is satisfied, $\max\left\{\|\overline{V}^k - \overline{V}^\star\|_\infty, \|\underline{V}^k - \underline{V}^\star\|_\infty\right\} \leqslant \epsilon$. This concludes our proof. □

In particular, the Bellman operator f and policy operator g^π are γ-contractive on $\mathbb{R}^S$, where γ is the discount factor. Therefore, Theorem 5 applies with $\alpha = \gamma$.

Remark 11. Theorem 6 implies that at the termination of Algorithm 1, the fixed point set $\mathcal{V}^\star$ can be *over-approximated* by

$$\mathcal{V}^\star \subseteq \mathcal{V}_{approx} := \prod_{s\in[S]} [\underline{V}_s^{k+1} - \epsilon, \overline{V}_s^{k+1} + \epsilon],$$

where k is the last iterate before Algorithm 1 terminates.

7. Path planning in non-stationary wind fields

We apply set-based value iteration to a wind-assisted probabilistic path planning problem in a non-stationary wind field (Wolf et al., 2010). MDP as a model for wind-assisted path planning of balloons in the stratosphere and exoplanets has recently gained traction (Bellemare et al., 2020; Wolf et al., 2010). Finite state–action MDPs have been shown to be a viable high-level path planning model (Wolf et al., 2010) for such applications.

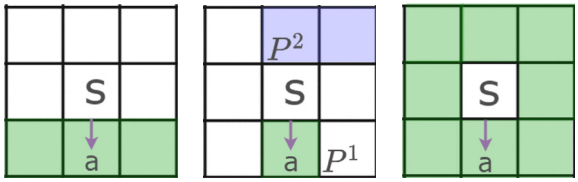
Mission Objective. In a two-dimensional wind-field, the wind-assisted balloon aims to reach target state $(8, 8)$ in Fig. 8 using minimum fuel.

Non-stationary Wind Fields. By collecting wind data on the environment's wind field, an MDP can be created and a policy that handles stochastic planning can be deployed. However, wind can be a time-varying factor that causes the *expected* optimal policy to have *worse-than-expected* worst-case performance. We built an ideal uncertain wind field to demonstrate how the set Bellman operator can be used to predict the best and worst-case behavior of a robust policy.

MDP Modeling Assumptions. Following Wolf et al. (2010), we model the path planning problem in a non-stationary wind field as an infinite horizon, discounted MDP with discrete state–actions. While balloons typically traverse in three dimensions, we assume that the wind is consistent in the vertical direction and that the final target is any vertical position along the given two-dimensional coordinates.

States. States are grouped into three different regions based on their wind variability as shown in Fig. 8: $[S] = [S_{calm}] \bigcup [S_{gusty}] \bigcup [S_{unreliable}]$.

(1) $\mathbf{s} \in [\mathbf{S_{calm}}]$. The wind magnitude changes between $[0, 0.5]$, and the wind direction changes between $[0, 2\pi]$. $[S_{calm}] = \{(i, j) \mid (0, 0) \leqslant (i, j) \leqslant (2, 8),\ (0, 6) \leqslant (i, j) \leqslant (8, 8)\}$.



(a) State transition in calm wind.

(b) State transition in unreliable wind.

(c) State transition in gusty wind.

Fig. 9. Transition probabilities for the three different wind regions.

(2) $\mathbf{s} \in [\mathbf{S_{gusty}}]$. The wind magnitude is 1, while the wind direction changes between $[0, 2\pi]$. $S_{gusty} = \{(i, j) \mid (3, 3) \leqslant (i, j) \leqslant (5, 5)\}$.

(3) $\mathbf{s} \in [\mathbf{S_{unreliable}}]$. A wind front occasionally moves across an otherwise windless region. The wind magnitude is either 0 or 1 and the wind direction changes between $[\pi/4, \pi/2]$.

Actions. The balloon is equipped with an actuator that provides a constant thrust of 1 unit power in 8 discretized directions shown in Fig. 8b. We assume that this actuation force is enough to move the balloon across one state in wind with magnitude $\leqslant 0.5$, and is otherwise not strong enough to overcome the wind effects. In addition, each state has an action corresponding to no thrust.

Transition Probabilities. The transition probabilities are region-dependent. In the states $[S_{calm}]$ and $[S_{gusty}]$, the transition dynamics are stochastic but stationary in time. In the states $[S_{unreliable}]$, the transition dynamics are stochastic but change over time. We define the following neighboring states for each state $s \in [S]$.

(1) $\mathcal{N}(s)$: all 8 neighboring states of state s.
(2) $\mathcal{N}(s, a, 0)$: the neighboring state of s in the direction of a.
(3) $\mathcal{N}(s, a, 1)$: the neighboring state of s in the direction of a plus the two states adjacent to the neighbor state, shown in green in Fig. 9a.
(4) $\mathcal{N}(s, a, 2)$: state s's neighboring state in the opposite direction of the action a plus its clockwise neighbor, shown in purple in Fig. 9b.

In the calm wind region, the transition probabilities are given by

$$P_{sa,s'} = \begin{cases} \frac{1}{\mathcal{N}(s,a,1)}, & s' \in \mathcal{N}(s, a, 1) \\ 0 & \text{otherwise,} \end{cases} \quad \forall\ s \in [S_{calm}]. \tag{57}$$

In the gusty wind region, the transition probabilities are given by

$$P_{sa,s'} = \begin{cases} \frac{1}{\mathcal{N}(s)}, & s' \in \mathcal{N}(s) \\ 0 & \text{otherwise,} \end{cases} \quad \forall\ s \in [S_{gusty}],\ \forall\ a \in [A]. \tag{58}$$

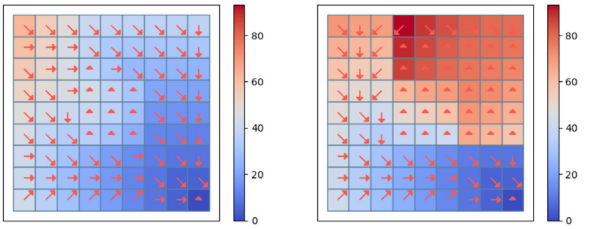
In the unreliable wind region, the transition probabilities vary between transition dynamics P_s^1 and P_s^2.

$$P^1_{sa,s'} = \begin{cases} 1, & s' \in \mathcal{N}(s, a, 0) \\ 0 & \text{otherwise,} \end{cases} \quad \forall\ s \in [S_{gusty}],\ \forall\ a \in [A]. \tag{59}$$

$$P^2_{sa,s'} = \begin{cases} 0.5, & s' \in \mathcal{N}(s, a, 2) \\ 0 & \text{otherwise,} \end{cases} \quad \forall\ s \in [S_{gusty}],\ \forall\ a \in [A]. \tag{60}$$

Collectively, P_s^1 and P_s^2 collectively form the uncertainty set $\mathcal{P}_s \subset \Delta_S^A$ defined at each state.

$$\mathcal{P}_s = \{P^i_{sa} \mid i \in \{1, 2\}, a \in [A]\},\ \forall s \in [S_{unreliable}]. \tag{61}$$



(a) Optimistic case with expected objective of 54.2

(b) Robust case with expected objective of 96.7.

Fig. 10. Visualization of optimistic and robust policies.

Table 1
Bellman, optimistic policy, robust policy value bounds.

Set	Maximum value	Minimum value
$\mathcal{V}^B$	70.61	62.25
$\mathcal{V}^o$	101.58	62.25
$\mathcal{V}^r$	70.63	70.52

Cost. At each state–action, the cost is the sum of the current distance from target position $s_{targ} = (8, 8)$, as well as the fuel expended by the given action.

$$C((i,j),a) = \sqrt{(i - s_{targ}[0])^2 + (j - s_{targ}[1])^2} + \frac{1}{2}\|a\|_2.$$

We take $\|a\|_2 = 1$ for all actions except for the staying still action, where $\|a\|_2 = 0$.

7.1. Bellman, optimistic policy, and robust policy

We compute the optimistic and robust bounds with parameter uncertainty in $\mathcal{P}$ when $s \in [S_{unreliable}]$ by running Algorithm 1. The results are shown in Fig. 10.

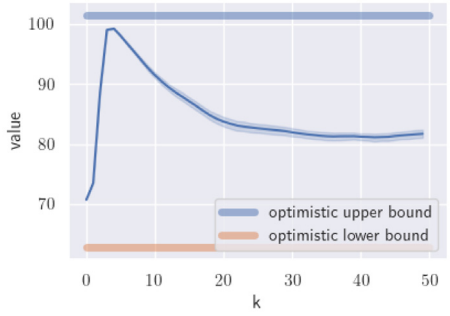
We denote the optimistic policy as π^o and the robust policy as π^r, and derive the bounds of their respective value vector sets $\mathcal{V}^o$ (38) and $\mathcal{V}^r$ (39) using Algorithm 1. The output is compared against the bounds of the set-based Bellman operator's fixed point set $\mathcal{V}^\star$ in Table 1.

Non-stationary wind field Next, we consider a non-stationary wind field: at each time step k, the transition probability P^k is chosen at random from $\mathcal{P}$ (61). We compare three different policy update schemes: (1) stationary optimistic policy π^o as policy operator g^o (38), (2) stationary robust policy π^r as policy operator g^r (39), and (3) Bellman policy that is one-step optimal for the MDP $([S], [A], P^k, C, \gamma)$ as f (9). These three different policy schemes are given by

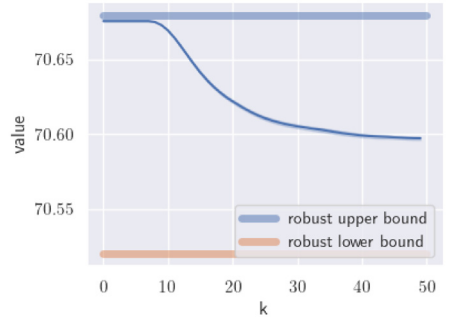
$$V^{k+1} = g^o(V^k, C, P^k), \tag{62}$$
$$V^{k+1} = g^r(V^k, C, P^k), \tag{63}$$
$$V^{k+1} = f(V^k, C, P^k). \tag{64}$$

The cost-to-go at state $s_{orig} = [0, 0]$ is plotted in Fig. 11. The optimistic policy (62) has the greatest variation in value over the course of 50 MDP time steps. Both the robust policy (63) and the Bellman policy (64) achieve better upper-bound at each MDP iteration. The Bellman policy (64) achieves less than 70 in cost-to-go on average, which is the best among all three policy schemes. As we discussed in Remark 9, the robust policy has the smallest variance in value in the presence of wind uncertainty, achieving a value difference of less than 0.1.

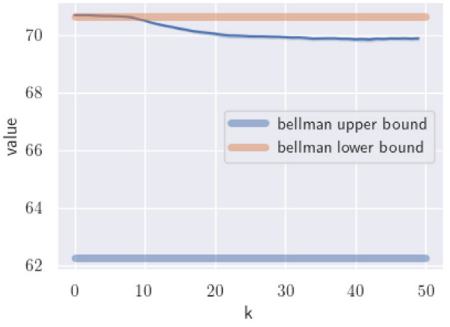
Sampled solutions. We can compute a sampled MDP model based on 50 samples of wind vectors for each state. Based on these samples, we add the action vector and compute the statistical distribution of state transitions. We then compute the value of these *stationary* sampled MDPs, and compare 9 randomly selected states' values. The resulting scatter plot is shown in Fig. 12.



(a) Optimistic Policy with $\mathcal{V}^o$'s bounding values.



(b) Robust Policy with $\mathcal{V}^r$'s bounding values.



(c) Bellman policy with $\mathcal{V}^B$'s bounding values.

Fig. 11. Comparison of robust policy, optimistic policy, and Bellman policy's value trajectories in non-stationary wind fields. Center blue line is the average over 50 trials. The shaded blue region denotes the standard variation. The top and bottom lines are the extrema values of the fixed points.

8. Conclusion

In this paper, we lifted contraction operators that solve Markov decision processes to operate on *compact sets* of vectors. Using fixed point analysis, we showed that the set-based value operators have fixed point sets that are invariant to the parameter uncertainties. These sets were applied to non-stationary and parameter uncertain MDPs to derive novel results in these

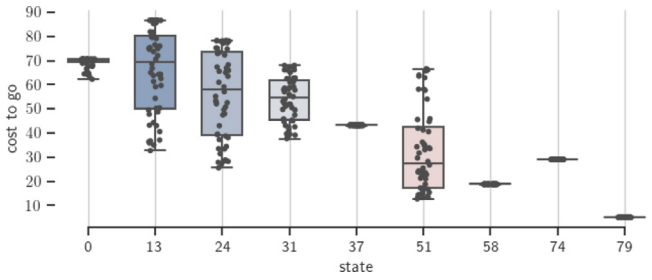


Fig. 12. Comparison of different optimal value vectors under the Bellman operator for 50 randomly sampled MDPs. On the x-axis, the state number is computed as $i \times 9 + j$.

settings. Finally, we demonstrated our results on path planning in a non-stationary wind field. In the future, we aim on apply set-based value operators to stochastic games in the presence of uncoordinated players.

Appendix A. Set sequence convergence

Lemma 5. *Let $\{\mathcal{V}_n\} \subseteq \mathcal{K}(\mathbb{R}^S)$ be a converging sequence for $d_{\mathcal{K}}$ with $\mathcal{V}_n \to \mathcal{V}$ as $n \to \infty$. For all $V \in \mathcal{V}$, there exists a converging subsequence $\{V^{\varphi(n)}\}_{n\in\mathbb{N}}$ whose limit is V for $\|\cdot\|_\infty$.*

Proof. Let $V \in \mathcal{V}$. We define the strictly increasing function φ on $\mathbb{N}$ as follows: $\varphi(0) := 0$ and for all $n \in \mathbb{N}$, $\varphi(n+1) := \min\{j > \varphi(n) \mid \exists V^j \in \mathcal{V}^j, \|V - V^j\|_\infty = d(V, \mathcal{V}^j) \leqslant (n+1)^{-1}\}$. Finally, as for all $n \in \mathbb{N}^*$, $\|V - V^{\varphi(n)}\|_\infty \leqslant (\varphi(n)+1)^{-1}$, the result holds. □

Appendix B. Proof of Lemma 1

Proof. Let $(V, m) \in \mathbb{R}^S \times \mathcal{M}$ and consider a sequence $\{(V_k, m_k)\}_{k\in\mathbb{N}} \subset \mathbb{R}^S \times \mathcal{M}$ that converges to (V, m). It holds that $\|h(V_k, m_k) - h(V, m)\|_\infty \leqslant \|h(V_k, m_k) - h(V, m_k)\|_\infty + \|h(V, m_k) - h(V, m)\|_\infty$, where from the α-contractive property of $h(\cdot, m^k)$, $\|h(V_k, m_k) - h(V, m_k)\|_\infty \leqslant \alpha\|V_k - V\|_\infty$. From the $K(V)$-Lipschitz property of $h(V, \cdot)$,

$$\|h(V, m_k) - h(V, m)\|_\infty \leqslant K(V)\|m_k - m\|_\infty.$$

As both $\lim_{k\to\infty} \|V_k - V\|_\infty \to 0$ and $\lim_{k\to\infty} \|m_k - m\|_\infty \to 0$, $\|h(V_k, m_k) - h(V, m)\|_\infty \to 0$ and h is continuous. □

Appendix C. Proof of Lemma 2

Proof. We show that both the Bellman operator f and the policy evaluation operator g^π satisfy the contracting/order preserving/Lipschitz properties given in Definition 4. Contraction: given $(C, P) \in \mathcal{M}$, $g^\pi(\cdot, C, P)$ and $f(\cdot, C, P)$ are both γ-contractions (Puterman, 2014, Prop. 6.2.4) on the complete metric space $(\mathbb{R}^S, \|\cdot\|_\infty)$, when $\gamma < 1$.

Order preservation: given $(C, P) \in \mathcal{M}$, the operator $g^\pi(\cdot, C, P)$ is order preserving (Puterman, 2014, Lem. 6.1.2). Consider $U, V \in \mathbb{R}^S$ where $U \leqslant V$. If $g^\pi(\cdot, C, P)$ is order-preserving, $g^\pi(U, C, P) \leqslant g^\pi(V, C, P)$ for all $\pi \in \Pi$. Taking the infimum over Π, we have $f(U, C, P) = \inf_{\pi\in\Pi} g^\pi(U, C, P) \leqslant \inf_{\pi\in\Pi} g^\pi(V, C, P) = f(V, C, P)$.

$K(V)$-Lipschitz: given $(C, P), (C', P') \in \mathcal{M}$ and $V \in \mathbb{R}^S$, we prove the following for each $s \in [S]$,

$$|f_s(V, C', P') - f_s(V, C, P)| \leqslant \|c'_s - c_s\|_\infty + \gamma\|P'_s - P_s\|_\infty \max\{\|\pi_s^\star\|_\infty, \|\hat{\pi}_s\|_\infty\}\|V\|_\infty. \quad \text{(C.1)}$$

We prove (C.1) by case: (1) $f_s(V, C', P') \geqslant f_s(V, C, P)$, and (2) $f_s(V, C', P') \leqslant f_s(V, C, P)$. For case (1), let $\hat{\pi}$ (10) be the optimal policy for $f(V, C', P')$ and $\pi^\star$ be the optimal policy for $f(V, C, P)$. For $s \in [S]$, suppose $f_s(V, C', P') \geqslant f_s(V, C, P)$, then $0 \leqslant f_s(V, C', P') - f_s(V, C, P) \leqslant (c'_s)^\top \hat{\pi}_s - c_s^\top \pi_s^\star + \gamma(P'_s\hat{\pi}_s)^\top V - \gamma(P_s\pi_s^\star)^\top V$. Since $\pi^\star$ is sub-optimal for $f(V, C', P')$, we upper bound $|f_s(V, C', P') - f_s(V, C, P)| \leqslant (c'_s - c_s)^\top \pi_s^\star + \gamma[(P'_s - P_s)\pi_s^\star]^\top V$. Since $\pi_s^\star, \hat{\pi}_s \in \Delta_A$, $\|\pi_s^\star\|_\infty \leqslant 1$. We conclude that (C.1) holds when $f_s(V, C', P') \geqslant f_s(V, C, P)$. For case (2), $f_s(V, C', P') \leqslant f_s(V, C, P)$, (C.1) also holds by similar arguments.

Letting $m' = (C', P')$ and $m = (C, P)$, we can upper bound $f(V, m) - f(V, m') = f - f'$ as

$$\|f - f'\|_\infty \leqslant \max_{s\in[S]}\{\|c'_s - c_s\|_\infty + \gamma\|(P_s - P'_s)^\top V\|_\infty\} \quad \text{(C.2)}$$

$$\leqslant \max(1, \gamma\|V\|_\infty)\|m - m'\|_\infty. \quad \text{(C.3)}$$

The policy evaluation operator g^π satisfies (C.1) if $\max\{\|\pi_s^\star\|_\infty, \|\hat{\pi}_s\|_\infty\}$ is replaced by $\|\pi_s\|_\infty$. Since $\|\pi_s\|_\infty \leqslant 1$, g^π is $K(V)$-Lipschitz. □

References

Al-Sabban, Wesam H., Gonzalez, Luis F., & Smith, Ryan N. (2013). Wind-energy based path planning for unmanned aerial vehicles using markov decision processes. In *2013 IEEE international conference on robotics and automation* (pp. 784–789). IEEE.

Bellemare, Marc G, Candido, Salvatore, Castro, Pablo Samuel, Gong, Jun, Machado, Marlos C, Moitra, Subhodeep, et al. (2020). Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, *588*, 77–82.

Bellemare, Marc G, Ostrovski, Georg, Guez, Arthur, Thomas, Philip, & Munos, Rémi (2016). Increasing the action gap: New operators for reinforcement learning. *Vol. 30*, In *Proceedings of the AAAI conference on artificial intelligence*.

Chen, Zaiwei, Maguluri, Siva Theja, Shakkottai, Sanjay, & Shanmugam, Karthikeyan (2021). Finite-sample analysis of off-policy TD-learning via generalized bellman operators. *Advances in Neural Information Processing Systems*, *34*, 21440–21452.

Doshi, Prashant, Goodwin, Richard, Akkiraju, Rama, & Verma, Kunal (2005). Dynamic workflow composition: Using markov decision processes. *International Journal of Web Services Research (IJWSR)*, *2*(1), 1–17.

Givan, Robert, Leach, Sonia, & Dean, Thomas (2000). Bounded-parameter Markov decision processes. *Artificial Intelligence*, *122*(1–2), 71–109.

Goyal, Vineet, & Grand-Clement, Julien (2022). Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*.

Henrikson, Jeff (1999). Completeness and total boundedness of the hausdorff metric. In *MIT undergraduate journal of mathematics*. Citeseer.

Iyengar, Garud N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, *30*(2), 257–280.

Kumar, Aviral, Zhou, Aurick, Tucker, George, & Levine, Sergey (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, *33*, 1179–1191.

Lecarpentier, Erwan, & Rachelson, Emmanuel (2019). Non-stationary Markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in Neural Information Processing Systems*, *32*.

Li, Sarah H. Q., Adjé, Assalé, Garoche, Pierre-Loïc, & Açıkmeşe, Behçet (2021). Bounding fixed points of set-based bellman operator and Nash equilibria of stochastic games. *Automatica*, *130*, Article 109685.

Mannor, Shie, Mebel, Ofir, & Xu, Huan (2016). Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, *41*(4), 1484–1509.

Melo, Francisco S. (2001). *Convergence of Q-learning: A simple proof*: *Tech. Rep*, (pp. 1–4). Institute of Systems and Robotics.

Neumann, J. v (1928). Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, *100*(1), 295–320.

Nilim, Arnab, & El Ghaoui, Laurent (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, *53*(5), 780–798.

Padakandla, Sindhu, KJ, Prabuchandran, & Bhatnagar, Shalabh (2020). Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *50*(11), 3590–3606.

Puterman, Martin L. (2014). *Markov decision processes.: Discrete stochastic dynamic programming*. John Wiley & Sons.

Rudin, Walter, et al. (1964). *Vol. 3, Principles of mathematical analysis*. McGraw-hill New York.

Schröder, Bernd S. W. (2003). Ordered sets. *Springer*, *29*, 30.

Van Hoof, Herke, Hermans, Tucker, Neumann, Gerhard, & Peters, Jan (2015). Learning robot in-hand manipulation with tactile features. In *2015 IEEE-RAS 15th international conference on humanoid robots (humanoids)* (pp. 121–127). IEEE.

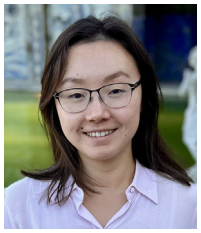
Wiesemann, Wolfram, Kuhn, Daniel, & Rustem, Berç (2013). Robust Markov decision processes. *Mathematics of Operations Research*, *38*(1), 153–183.

Wolf, Michael T, Blackmore, Lars, Kuwata, Yoshiaki, Fathpour, Nanaz, Elfes, Alberto, & Newman, Claire (2010). Probabilistic motion planning of balloons in strong, uncertain wind fields. In *2010 IEEE international conference on robotics and automation* (pp. 1123–1129). IEEE.

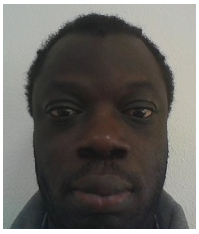
Xu, Huan, & Mannor, Shie (2010). Distributionally robust Markov decision processes. *Advances in Neural Information Processing Systems*, *23*.

Yang, Insoon (2017). A convex optimization approach to distributionally robust Markov decision processes with wasserstein distance. *IEEE Control Systems Letters*, *1*(1), 164–169.

Yu, Pengqian, & Xu, Huan (2015). Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, *61*(9), 2538–2543.



Sarah H.Q. Li is a postdoctoral scholar at ETH Zurich. She received her Ph.D. in Aeronautics and Astronautics Engineering at the University of Washington and her B.A.Sc degree in Engineering Physics at the University of British Columbia in 2017. She was a visiting researcher at Onera, the French Aerospace Lab in 2019 and has interned for Google Loon (2020), Deutsche Elektronen-Synchrotron (2015), and Macdonald Dettwiler and Associates (2014). Her research interests include game theory, Markov decision processes, and optimization.



Assalé Adjé obtained his PHD in applied mathematics/computer science from the École Polytechnique (Palaiseau, France) in 2011. Since 2016, he has been an associate professor in computer science at the University of Perpignan Via Domitia. His research interests include game theory and policy iteration algorithms. He is also interested in optimization theory and their applications in dynamical systems and programs verification.

Pierre-Loïc Garoche is a professor at ENAC, the French National School of Civil Aviation, and a contractor for NASA Ames in the Robust Software Engineering group. From 2008–2020, he worked at Onera, the French Aerospace Lab. His work is focused on the formal verification of control system software.

Behçet Açikmeşe is a professor of Aerospace Optimization and Control at University of Washington, Seattle. He received his Ph.D. in Aerospace Engineering from Purdue University.

Previously, he was a senior technologist at JPL, and faculty member at the University of Texas at Austin. At JPL, he developed flyaway control algorithms that were successfully used in the landing of Curiosity and Perseverance rovers on Mars. His research interests include robust and nonlinear control, convex optimization and its applications control theory and its aerospace applications, and Markov decision processes. He is a recipient of many NASA and JPL achievement awards for his contributions to spacecraft control in planetary landing, formation flying, and asteroid and comet sample return missions. He is also a recipient of NSF CAREER Award and IEEE CSS Award for Technical Excellence in Aerospace Control. He is a fellow of IEEE and AIAA.