Set-based Value Operators for Non-stationary and Uncertain Markov Decision Processes [*]

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

[a] *Department of Aeronautics and Astronautics, University of Washington, Seattle, USA. (e-mail:{sarahli, behcet}@uw.edu).*

[b] *LAMPS, Université de Perpignan Via Domitia, Perpignan, France. (e-mail: assale.adje@univ-perp.fr).*

[c] *École Nationale de l'Aviation Civile, Université de Toulouse, Toulouse, France. (e-mail: Pierre-Loic.Garoche@enac.fr).*

Abstract

This paper analyzes finite state Markov Decision Processes (MDPs) with uncertain parameters in compact sets and re-examines results from robust MDP via set-based fixed point theory. To this end, we generalize the Bellman and policy evaluation operators to contracting operators on the value function space and denote them as *value operators*. We lift these value operators to act on *sets* of value functions and denote them as *set-based value operators*. We prove that the set-based value operators are *contractions* in the space of compact value function sets. Leveraging insights from set theory, we generalize the rectangularity condition in classic robust MDP literature to a containment condition for all value operators, which is weaker and can be applied to a larger set of parameter-uncertain MDPs and contracting operators in dynamic programming. We prove that both the rectangularity condition and the containment condition sufficiently ensure that the set-based value operator's fixed point set contains its own extrema elements. For convex and compact sets of uncertain MDP parameters, we show equivalence between the classic robust value function and the supremum of the fixed point set of the set-based Bellman operator. Under dynamically changing MDP parameters in compact sets, we prove a set convergence result for value iteration, which otherwise may not converge to a single value function. Finally, we derive novel guarantees for probabilistic path planning problems in planet exploration and stratospheric station-keeping.

Key words: Markov decision process, contraction operator, stochastic control, decision making and autonomy

1 Introduction

Markov decision process (MDP) is a versatile model for decision making in stochastic environments and is widely used in trajectory planning [1], robotics [22], and operations research [4]. 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. Known as dynamic programming approaches, the resulting solutions extend to the dynamic-free setting via value-based reinforcement learning [20].

Many application settings of MDPs, including traffic light control, motion planning, and dexterous manipulation, deal with *environmental non-stationarity*—dynamically changing MDP cost and transition probabilities due to external factors or the presence of

interfering decision makers. This environmental non-stationarity corresponds to uncertainty in the MDP transition and cost parameters and differs from an MDP's internal stochasticity, which corresponds to stationary MDP transition and cost parameters and models the stochastic dynamics whose probability distributions do not change over time. Under environmental non-stationarity, robust MDP, risk-sensitive reinforcement learning, and zero-sum stochastic games derive policies for the optimal worst-case performance—value functions that result from adversarial selections of the MDP parameters. A standard approach is to take an appropriate Bellman operator-variant and solve for its fixed point using a minmax formulation. By assuming that the MDP parameters are chosen adversarially, the worst-case approach is able to derive asymptotic bounds of the value function trajectory affected by the parameter non-stationarity.

Much of the robust MDP literature focuses on the worst-case analysis of the *MDP policy improvement* and

---

[*] This paper was not presented at any IFAC meeting. Corresponding author Sarah H.Q. Li. Email. sarahli@uw.edu.

MDP policy evaluation problem under a rectangularity assumption on a set of MDP parameters. Mathematically, these problems correspond to the Bellman and policy evaluation operators. However, recent progress in learning-based methods utilizes other contraction operators such as Q-learning [23] and temporal difference [21]. In order to broaden the application of our results, we ground our analysis in a more general class of value operators, and ask the following question from a less adversarial perspective: is it possible to characterize the transient behavior of an MDP contraction operator with dynamically changing parameters?

Contributions. For compact sets of finite state MDP parameter uncertainties, we propose the set-extensions of *value operators*: a general class of contraction operators that are order-preserving on the space of value functions and Lipschitz in the space of MDP parameters. We prove the existence of compact *fixed point sets* of the set-based value operators and show that the set-based value iteration converges. In a non-stationary Markovian environment, standard value iteration may not converge. However, we can show that the point-to-set distance of the resulting value function trajectory to the fixed point set always goes to zero in the limit. We derive a *containment condition* that is sufficient for the fixed point sets to contain their own extremal elements. Within robust MDPs, we show that the containment condition generalizes the rectangularity condition, such that the optimal worst-case policy, or the robust policy, exists when the containment condition is satisfied. We then derive the relationship between the fixed point sets of 1) the set-based optimistic policy evaluation operator, 2) the set-based robust policy evaluation operator, and 3) the set-based Bellman operator. Given a value operator and a compact MDP parameter uncertainty set, we present an algorithm that computes the bounds of the corresponding fixed point set and derive its convergence guarantees. Finally, we apply our results to the wind-assisted navigation of high altitude platform systems relevant to space exploration [25] 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 [5, 8, 15, 24]. The cost-to-go under parameter uncertainty with coupling between states and time steps is similarly bounded in [6, 12]. The effect of statistical uncertainty on the optimal cost-to-go is studied in [12, 15, 24, 26]. Recently, MDP with parameter uncertainty has 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 [9, 10, 16]. Most RL research extends the minmax 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 [3].

As opposed to the worst-case approach to analyzing MDPs under parameter uncertainty, we do not assume adversarial MDP parameter selection. Instead, we derive a set of cost-to-gos that is invariant with respect to compact parameter uncertainty sets for order-preserving, α-contracting operators, a class that the Bellman operator belongs to. We continue from our previous work [11], in which we analyzed the set-based Bellman operator for cost uncertainty only.

Notation: A set of N elements is given by $[N] = \{0, \ldots, N-1\}$. 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 $\mathbf{1}_N = [1, \ldots, 1]^T \in \mathbb{R}^{N \times 1}$. 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 $(h_1, \ldots, h_s)$, where h_s is the value of h in the sth coordinate, $s \in [S]$. Throughout the paper, $\|\cdot\|$ denotes the infinity norm in $\mathbb{R}^S$.

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 the discount factor, $[S] = \{1, \ldots, S\}$ is the **finite set of states** and $[A] = \{1, \ldots, A\}$ is the **finite set of actions**. Without loss of generality, assume that each action is admissible from each state $s \in [S]$.

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 $c_s = [C_{s1}, \ldots, C_{sA}] \in \mathbb{R}^A$, such that $C = [c_1, \ldots, c_S]^\top$.

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 $P_s = [p_{s1}, \ldots, p_{sA}] \in \Delta_S^A \subset \mathbb{R}^{S \times A}$, and all possible transition probabilities in the MDP are given by the matrix $P = [P_1, \ldots, P_S] \in \Delta_S^{SA} \subset \mathbb{R}^{S \times SA}$.

MDP Objective. We want to minimize the expected cost-to-go, or the **value vector** $V \in \mathbb{R}^S$, defined per state as

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

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 value vector can be minimized by the choice of actions a^t at every time step t:

$$V_s^\star := \min_{a^t \in [A]} \mathbb{E}_s\left\{ \sum_{t=0}^\infty \gamma^t C_{s^t a^t} \mid s^0 = s \right\}, \ \forall \ s \in [S], \tag{3}$$

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.*

MDP Policy. We optimize the objective (3) via a *policy*, denoted as $\pi = [\pi_1, \ldots, \pi_S] \in \Delta_A^S$, where the ath element of $\pi_s \in \Delta_A$ is the conditional probability of action a being chosen from state s. 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 $P_s \pi_s \in \Delta_S$.

2.1 Value operators

Solving an MDP is equivalent to finding the value vector and the associated policy that minimizes the objective (3). Typical solution methods utilize *order preserving* [19, Def.3.1], *α-contractive operators* whose fixed points are the optimal value vectors (e.g. Bellman operator [17, Thm.6.2.3], Q-value operator [13]).

Definition 2 (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*

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

Definition 3 (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 MDP parameter space.

Definition 4 ($K(V)$-Lipschitz) *Let $(\mathcal{X}, d_\mathcal{X})$ be a metric space with metric $d_\mathcal{X}$ and $(\mathcal{Y}, d_Y)$ 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*

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

Remark 5 *The α-contraction property is a special instance of Lipschitz continuity in which the input and output spaces are identical and the Lipschitz constant is less than 1.*
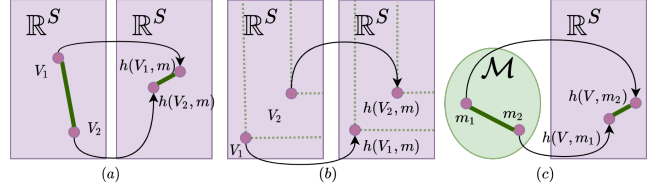

Fig. 1. Illustration of the three value operator properties. (a) α-contraction on $\mathbb{R}^S$, (b) Order preservation on $\mathbb{R}^S$, and (c) $K(V)$-Lipschitz in input space $\mathcal{M}$.

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

Definition 6 (Value operator) *Consider the operator h,*

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

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

(1) For all $m \in \mathcal{M}$, $h(\cdot, m)$ is an α-contraction in $\mathbb{R}^S$.
(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 7 *While we only consider value operators whose input's first component is $\mathbb{R}^S$, Definition 6 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 6 [13].*

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

Lemma 8 (Continuity) *If h (6) is a value operator on $\mathbb{R}^S \times \mathcal{M}$, h is continuous on $\mathbb{R}^S \times \mathcal{M}$.*

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 $\|h(V_k, m_k) - h(V, m)\| \leqslant \|h(V_k, m_k) - h(V, m_k)\| + \|h(V, m_k) - h(V, m)\|$, where from the α-contractive property of $h(\cdot, m^k)$, $\|h(V_k, m_k) - h(V, m_k)\| \leqslant \alpha \|V_k - V\|$. From the $K(V)$-Lipschitz property of $h(V, \cdot)$,

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

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

We make the following assumption on the MDP parameter set $\mathcal{M}$ with respect to h.

Assumption 9 (Containment condition) *The MDP parameter set $\mathcal{M}$ satisfies the containment condition with respect to h if $\mathcal{M}$ is compact and for all $V \in \mathbb{R}^S$,*

$$\bigcap_{s\in[S]} \underset{m\in\mathcal{M}}{\arg\min}\, h_s(V,m) \neq \varnothing, \quad \bigcap_{s\in[S]} \underset{m\in\mathcal{M}}{\arg\max}\, h_s(V,m) \neq \varnothing. \tag{7}$$
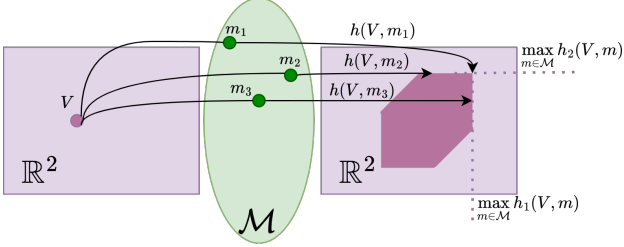

Fig. 2. We illustrate $\arg\max_{m\in\mathcal{M}} h_s(V,m)$ for a value operator h when $S = 2$. Here, $\arg\max_{m\in\mathcal{M}} h_1(V,m) = \{m_2, m_3\}$, $\arg\max_{m\in\mathcal{M}} h_2(V,m) = \{m_1, m_2\}$. Therefore, m_2 is the common parameter that achieves $\max_{m\in\mathcal{M}} h_s(V,m)$ for all $s \in [S]$.

Remark 10 *Assumption 9 is an h-dependent condition imposed on the structure of $\mathcal{M}$, and is independent of $\mathcal{M}$'s convexity and connectivity.*

2.2 Bellman and policy evaluation operators

Examples of value operators include the Bellman operator and the policy evaluation operators when the MDP cost and transition probability are input parameters rather than fixed parameters.

Definition 11 (Policy evaluation operator) *Given a policy $\pi \in \Pi$, 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*

$$g_s^\pi(V,C,P) := c_s^\top \pi_s + \gamma \left(P_s\pi_s\right)^\top V, \ \forall s \in [S]. \tag{8}$$

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 $V^\pi(C,P)$ [17, Thm.6.2.5].

$$V^\pi(C,P) = g^\pi(V^\pi, C, P), \ V^\pi(C,P) \in \mathbb{R}^S. \tag{9}$$

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

Definition 12 (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*

$$f_s(V,C,P) := \inf_{\pi_s\in\Delta_A} g_s^\pi(V,C,P), \ \forall s \in [S]. \tag{10}$$

The corresponding optimal policy $\pi^\star = (\pi_1^\star, \ldots, \pi_s^\star)$ is defined per state as $\pi_s^\star \in \arg\min_{\pi_s} g_s^\pi(V,C,P)$ (10) and satisfies $\mathbf{1}_S^\top \pi_s^\star = 1 \ \forall s \in [S]$. One such policy is defined for all $(s,a) \in [S] \times [A]$ by

$$\pi_{s,a}^\star := \begin{cases} > 0 & a \in \underset{a\in[A]}{\arg\min}\, C_{sa} + \gamma p_{sa}^\top V, \\ 0 & otherwise. \end{cases} \tag{11}$$

where $\arg\min_{a\in[A]}(h)$ is the set of minimizing actions for the function h. An optimal policy in the form (11) always exists for a discounted infinite horizon MDP [17, Thm 6.2.10]. Given 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 $V^B(C,P)$.

$$V^B(C,P) = f(V^B, C, P), \ V^B(C,P) \in \mathbb{R}^S. \tag{12}$$

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

We show that both (8) and (10) are value operators.

Lemma 13 *The Bellman operator (10) and the policy evaluation operators (8) for all $\pi \in \Pi$ are value operators on $\mathbb{R}^S \times \mathcal{M}$ where $\mathcal{M} \subseteq \mathbb{R}^{S\times A} \times \Delta_S^{SA}$ (6).*

PROOF. We show that both the Bellman operator f and the policy evaluation operator g^π satisfy the contractive, order preserving and Lipschitz properties given in Definition 6. Contraction: given $(C,P) \in \mathcal{M}$, $g^\pi(\cdot, C, P)$ and $f(\cdot, C, P)$ are both γ-contractions [17, Prop.6.2.4] on the complete metric space $(\mathbb{R}^S, \|\cdot\|_\infty)$, where $\gamma < 1$ is the discount factor.

Order preservation: given $(C,P) \in \mathcal{M}$, the operator $g^\pi(\cdot, C, P)$ is order preserving [17, 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 $V \in \mathbb{R}^S$, let $\hat{\pi}$ (11) 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 can upper bound $|f_s(V,C',P') - f_s(V,C,P)| \leqslant (c_s' - c_s)^\top \pi_s^\star + \gamma[(P_s' - P_s)\pi_s^\star]^\top V$. We conclude that

$$\begin{aligned} &|f_s(V,C',P') - f_s(V,C,P)| \\ &\leqslant \|c_s' - c_s\|_\infty + \gamma \|P_s' - P_s\|_\infty \max\{\|\pi_s^\star\|_\infty, \|\hat{\pi}_s\|_\infty\} \|V\|_\infty. \end{aligned} \tag{13}$$

4

Since $\pi_s^\star, \hat{\pi}_s \in \Delta_A$, $\|\pi_s^\star\|_\infty \leqslant 1$ and $\|\hat{\pi}_s\|_\infty \leqslant 1$. By similar arguments, (13) is true if $f_s(V, C', P') \leqslant f_s(V, C, P)$. We can upper bound $f(V, m) - f(V, m') = f - f'$ as

$$\|f - f'\|_\infty \leqslant \max_{s \in [S]} \{\|c_s' - c_s\|_\infty + \gamma \|(P_s - P_s')^\top V\|_\infty\} \tag{14}$$

$$\leqslant \max(1, \gamma \|V\|_\infty) \|m - m'\|_\infty. \tag{15}$$

The policy evaluation operator g^π satisfies (13) 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. $\square$

Remark 14 *Beyond the policy evaluation operator and the Bellman operator, many algorithms in reinforcement learning can be reformulated using value operators. For example, it's not difficult to show that the Q-learning operator [13] is a value operator on the vector space $\mathbb{R}^{SA}$.*

2.3 Containment-satisfying MDP parameter sets

Assumption 9 restricts the structure of $\mathcal{M}$ with respect to the value operator h. Thus whether or not $\mathcal{M}$ satisfies Assumption 9 must always be determined with respect to the operator h. With respect to the Bellman operator f (10) and the policy evaluation operators g^π (8), the following conditions in robust MDP are sufficient to satisfy Assumption 9.

Definition 15 ((s, a)-rectangular sets [8, 15]) *The uncertainty set $\mathcal{M} \subset \mathbb{R}^{S \times A} \times \Delta_S^{SA}$ is (s, a)-rectangular if*

$$\mathcal{M} = \bigtimes_{(s,a) \in [S] \times [A]} \mathcal{M}_{sa}, \mathcal{M}_{sa} \subset \mathbb{R} \times \Delta_S, \forall (s, a) \in [S] \times [A]. \tag{16}$$

Intuitively, (s, a)-rectangularity implies that the MDP parameter uncertainty is *decoupled* between each state-action. A more general condition is if the parameter uncertainty is decoupled between different states but not between different actions within the same state.

Definition 16 (s-rectangular sets) *The uncertainty set $\mathcal{M} \subset \mathbb{R}^{S \times A} \times \Delta_S^{SA}$ is s-rectangular if*

$$\mathcal{M} = \bigtimes_{s \in [S]} \mathcal{M}_s, \mathcal{M}_s \subset \mathbb{R}^A \times \Delta_S^A, \forall s \in [S]. \tag{17}$$

s-rectangularity generalizes (s, a)-rectangularity—i.e. (s, a)-rectangularity implies s-rectangularity.

Example 17 (Wind uncertainty) *Consider an MDP in which the states correspond to geographical coordinates, the actions correspond to navigation choices (up, down, left, right), and the transition probabilities*
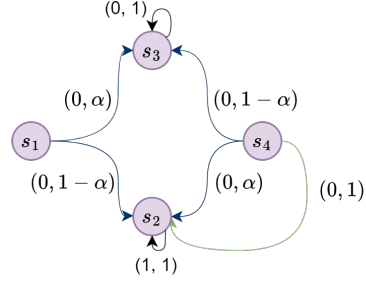

Fig. 3. MDP with parameter coupling in transition probability across different states.

correspond to the local wind patterns that vary between N major wind trends over time per state. At state $s \in [S]$, the transition probabilities of trend $i \in [N]$ are given by P_s^i. If the wind pattern strictly switches between the discrete wind trends, then the transition uncertainty at state $s \in [S]$ is $\mathcal{P}_s = \{P_s^1, \ldots, P_s^N\}$. If the wind pattern is a mixture of the 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 $\mathcal{P} = \times_{s \in [S]} \mathcal{P}_s$.

We show that the rectangularity conditions indeed are sufficient for satisfying Assumption 9 with respect to f (10) and g^π (8).

Proposition 18 *If $\mathcal{M}$ is compact and s-rectangular (Definition 16), $\mathcal{M}$ satisfies Assumption 9 with respect to f (10) and g^π (8) for all $\pi \in \Pi$.*

PROOF. We first show that $\mathcal{M}$ satisfies Assumption 9 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 8, 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 $(C^\star, P^\star) \in \operatorname{argmin}_{m \in \mathcal{M}} f_s(V, C, P)$ for all $s \in [S]$. We conclude that $\mathcal{M}$ satisfies Assumption 9.

Given $\pi \in \Pi$ 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 $(C^\star, P^\star) \in \mathcal{M}$. $\square$

Beyond s-rectangularity, there are sets that satisfy Assumption 9 with respect to specific value operators.

Example 19 (Beyond rectangularity) *In Figure 3, we visualize a four state MDP with transition uncertainty $\mathcal{M}$ parameterized by α. MDP states are the nodes and MDP actions are the arrows. Actions that transition to multiple states are visualized by multi-headed arrows.*

Each head has an associated tuple $(c_{sa}, p_{sa,s'})$ denoting its state-action cost and transition probability. All states have a single action except for state s_4, where two actions exist and are distinguished by different colors. Both s_2 and s_3 are absorbing states with a unique action, such that $V_2 = \frac{1}{1-\gamma}$ and $V_3 = 0$ for both f and g^π for all $\pi \in \Pi$, where γ is the discount factor.

The states s_1 and s_4 have transition uncertainty parametrized by $\alpha \in [0,1]$. Therefore, $\mathcal{M}$ violates s-rectangularity (Definition 16). The optimal cost-to-go values V_1 and V_4 occur at different α's. Therefore, $\mathcal{M}$ violates Assumption 9 with respect to f. However, suppose that at s_4, we only consider policies that exclusively choose the action colored green in Fig. 3. Then the expected cost-to-go at s_4, V_4, is independent of α. The minimum and maximum values of V_1 under π occur at $\alpha = 1$ and $\alpha = 0$, respectively. Therefore, $\mathcal{M}$ satisfies Assumption 9 with respect to operator g^π for all $\pi = [\pi_{s_1}, \ldots, \pi_{s_4}]$ where $\pi_{s_4} = [1,0]$.

3 Set-based value operators

Motivated by the uncertain MDP parameters encountered in robust MDP, stochastic games, and reinforcement learning in uncertain environments, we now consider value operators with respect to a compact set of uncertain MDP parameters. To understand the effect of both stationary and dynamic parameter uncertainty on the value vector, we extend value operators to set-based value operators, and prove the existence of fixed point sets on the space of compact subsets of $\mathbb{R}^S$.

To facilitate our set-based analysis, we first introduce Hausdorff-type set distances.

Definition 20 (Point-to-set Distance) *The distance between a value vector and a set $\mathcal{V} \subseteq \mathbb{R}^S$ is given by*

$$W \mapsto d(W, \mathcal{V}) := \inf_{V \in \mathcal{V}} \|W - V\|. \qquad (18)$$

On the space of compact subsets of $\mathbb{R}^S$, given by $\mathcal{K}(\mathbb{R}^S)$, the distance between value vector sets extends (18) and is given by the Hausdorff distance [7].

Definition 21 (Set-to-set Distance) *The Hausdorff distance between two value vector sets $\mathcal{V}, \mathcal{W} \subseteq \mathbb{R}^S$ is given by*

$$d_{\mathcal{K}}(\mathcal{V}, \mathcal{W}) := \max\left\{\sup_{V \in \mathcal{V}} d(V, \mathcal{W}), \sup_{W \in \mathcal{W}} d(W, \mathcal{V})\right\}. \quad (19)$$

We use $(\mathcal{K}(\mathbb{R}^S), d_{\mathcal{K}})$ to denote the metric space formed by the set of all compact subsets of $\mathbb{R}^S$ under the Haus-
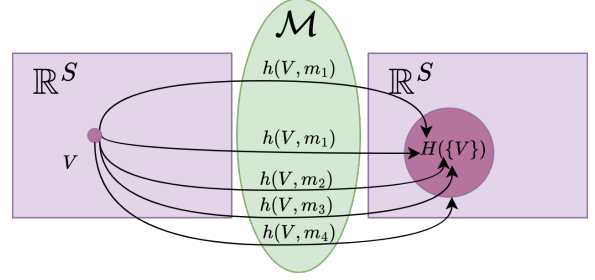

Fig. 4. Illustration of the set-based operator $H(\mathcal{V})$ applied to the singleton set $\mathcal{V} = \{V\} \subset \mathbb{R}^S$, we compute $h(V, m)$ for every parameter $m \in \mathcal{M}$ and collect the output $h(V, m)$, such that $H(\mathcal{V}) = \cup_{m \in \mathcal{M}} h(V, m)$.

dorff distance $d_{\mathcal{K}}$. The induced Hausdorff space is complete if and only if the original metric space is complete [7, Thm 3.3]. Therefore, $(\mathcal{K}(\mathbb{R}^S), d_{\mathcal{K}})$ is a complete metric space.

For a value operator h (6), we ask the following question: what is the *set* of possible value vectors when the MDP has parameter uncertainty given by $\mathcal{M}$? To resolve this, we define the set-based value operator H.

Definition 22 (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*

$$H(\mathcal{V}) := \{h(V, m) \mid (V, m) \in \mathcal{V} \times \mathcal{M}\} \subseteq \mathbb{R}^S, \quad (20)$$

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

We denote the set-based value operator induced by the Bellman operator (10) and policy evaluation operators (8) as F and G^π, respectively, such that for any value vector set $\mathcal{V} \subseteq \mathbb{R}^S$,

$$F(\mathcal{V}) := \{f(V, C, P) \mid (V, C, P) \in \mathcal{V} \times \mathcal{M}\}, \quad (21)$$

$$G^\pi(\mathcal{V}) := \{g^\pi(V, C, P) \mid (V, C, P) \in \mathcal{V} \times \mathcal{M}\}, \; \forall \, \pi \in \Pi. \quad (22)$$

The set-based Bellman operator F is the union over all the *optimal* value vectors, where the optimal policy that corresponds to each $f(V, C, P) \in F(\mathcal{V})$ varies based on (C, P). On the other hand, G_π is the union over all value vectors that results from a constant π and all possible $(C, P) \in \mathcal{M}$ parameters.

We can ask the following question: is there a set of value vectors that is invariant with respect to H? Similar to the value operators h from Definition 6, we can affirmatively answer this question by showing that H is α-contractive on $\mathcal{K}(\mathbb{R}^S)$.

Theorem 23 *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 (20) satisfies*

(1) For all $\mathcal{V} \in \mathcal{K}(\mathbb{R}^S)$, $H(\mathcal{V}) \in \mathcal{K}(\mathbb{R}^S)$;
(2) H is an α-contractive on $(\mathcal{K}(\mathbb{R}^S), d_\mathcal{K})$ (19) with a unique fixed point set $\mathcal{V}^\star$ given by

$$H(\mathcal{V}^\star) = \mathcal{V}^\star, \quad \mathcal{V}^\star \in \mathcal{K}(\mathbb{R}^S); \qquad (23)$$

(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 (21) and G^π (22), whose fixed point sets are denoted as $\mathcal{V}^B$ and $\mathcal{V}^\pi$, respectively.

$$F(\mathcal{V}^B) = \mathcal{V}^B \in \mathcal{K}(\mathbb{R}^S), \; G^\pi(\mathcal{V}^\pi) = \mathcal{V}^\pi \in \mathcal{K}(\mathbb{R}^S), \; \forall \pi \in \Pi. \tag{24}$$

PROOF. The first statement follows from Lemma 8, since the image of a compact set by a continuous function is compact [18]. Let us prove the second statement: for some $\beta \in (0, 1)$, for all $\mathcal{V}, \mathcal{V}' \in \mathcal{K}(\mathbb{R}^S)$:

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

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

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

We use the same technique to prove that

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

Finally, $d_\mathcal{K}(H(\mathcal{V}), H(\mathcal{V}')) \leqslant \alpha d_\mathcal{K}(\mathcal{V}, \mathcal{V}')$. From the Banach fixed point theorem and the completeness of $(\mathcal{K}(\mathbb{R}^S), d_\mathcal{K})$ [7, Thm 3.3], H has a unique fixed point $H^\star$ in $\mathcal{K}(\mathbb{R}^S)$.

The third point is a consequence of the Banach fixed point theorem. Finally, f and g^π are value operators (6) on $\mathbb{R}^S \times \mathcal{M}$, therefore this theorem's statements apply. $\square$

Remark 24 (Set-based value iteration) *An important consequence of Theorem 23 is the existence of the set-based value iteration, given by*

$$\mathcal{V}^{k+1} = H(\mathcal{V}^k), \; \mathcal{V}^0 \in \mathcal{K}(\mathbb{R}^S). \tag{25}$$

Analogous to standard value iteration, (25) is a sequence of value vector sets in $\mathcal{K}(\mathbb{R}^S)$ that converges to the fixed point set $\mathcal{V}^\star \in \mathcal{K}(\mathbb{R}^S)$.

4 Properties of the fixed point set

For the MDP parameters (C, P), the fixed point of $h(\cdot, C, P)$ is typically meaningful for the corresponding MDP. For example, the fixed point of a policy evaluation operator $g^\pi(\cdot, C, P)$ (8) is the expected cost-to-go under policy π, and the fixed point of the Bellman operator $f(\cdot, C, P)$ (10) is the minimum cost-to-go when π can be freely chosen. In this section, we derive properties of the fixed point set $\mathcal{V}$ of H (20) in the context of non-stationary value iteration.

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 discussed in [15], where at every iteration, a new set of MDP parameters m^k is chosen from $\mathcal{M}$ as

$$V^{k+1} = h(V^k, m^k), \; V^0 \in \mathbb{R}^S, \; m^k \in \mathcal{M}, \forall k \in \mathbb{N}. \tag{26}$$

In robust MDP literature [8, 15], m^k is modified by an adversarial opponent of the MDP decision maker such that (26) converges to a worst-case value vector. We consider a more general scenario in which m^k is chosen from the closed and bounded set $\mathcal{M}$ without any probabilistic prior. 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 (20).

Proposition 25 *Let $\mathcal{V}^\star$ be the fixed point set of the set-based value operator H (20) induced by h on $\mathbb{R}^S \times \mathcal{M}$ (6). If the non-stationary value iteration (26) satisfies $\{m^k\}_{k\in\mathbb{N}} \subset \mathcal{M}$, then the sequence $\{V^k\}_{k\in\mathbb{N}}$ defined by (26) 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 $\{V^k\}_{k\in\mathbb{N}}$ be a sequence defined by $\mathcal{V}^0 = \{V^0\}$ and $\mathcal{V}^{k+1} = H(\mathcal{V}^k)$, where H (20) is the set operator induced by h on $\mathbb{R}^S \times \mathcal{M}$. We first show statement 1). From Theorem 23, $\lim_{k\to\infty} \mathcal{V}^k$ converges to $\mathcal{V}^\star$ in $d_\mathcal{K}$. Therefore, $0 \leqslant d(V^k, \mathcal{V}^\star) = \inf_{y\in\mathcal{V}^\star} \|V^k - y\|_\infty \leqslant \sup_{x\in\mathcal{V}^k} \inf_{y\in\mathcal{V}^\star} \|x - y\|_\infty \leqslant d_H(\mathcal{V}^k, \mathcal{V}^\star) \to 0$ 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 $k \neq 0$, $\psi_1(k) := \min\{N > \psi_1(k-1) : \forall n \geqslant N, d(V^n, \mathcal{V}^\star) < (k+1)^{-1}\}$. Then, for all $k \in \mathbb{N}^\star$, there exists $y^{\psi_1(k)} \in \mathcal{V}^\star$ such that $\left\| V^{\psi_1(k)} - y^{\psi_1(k)} \right\| < (k+1)^{-1}$. 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 $y^\star \in \mathcal{V}^\star$ [18, 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 $l' \geqslant K_2$, $\left\| y^{\psi_1(\psi_2(l'))} - y^\star \right\| < \varepsilon/2$. So, taking $k \geqslant \max\{K_1, K_2\}$, we have $\left\| V^{\psi_1(\psi_2(k))} - y^\star \right\| \leqslant \left\| V^{\psi_1(\psi_2(k))} - y^{\psi_1(\psi_2(k))} \right\| + \left\| y^{\psi_1(\psi_2(k))} - y^\star \right\| \leqslant \varepsilon$ and $(V^{\psi_1(\psi_2(k))})_k$ converges to $y^\star \in \mathcal{V}^\star$.

$\square$

In addition to containing all asymptotic behavior of value vector trajectories under time-varying value iteration, the fixed point set $\mathcal{V}$ also contains all fixed points of the value operator $h(\cdot, C, P)$ when $(C, P) \in \mathcal{M}$ (6) are fixed.

Corollary 26 *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 (20), $V \in \mathcal{V}^\star$.*

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 25, $V \in \mathcal{V}^\star$ follows. $\square$

Going further, we can bound the transient behavior of (26) when V^0 is an element of the fixed point set $\mathcal{V}^\star$.

Corollary 27 (Transient behavior) *Let $\mathcal{V}^\star$ be the fixed point of the set-based value operator H (20) 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 (26) satisfies $\{V^k\}_{k \in \mathbb{N}} \subseteq \mathcal{V}^\star$.*

PROOF. As a fixed point set of H (20), $\mathcal{V}^\star$ (23) 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. $\square$

Remark 28 *Proposition 25 and Corollary 27 bound the asymptotic and transient behavior of the sequence $\{h(V^k, m^k)\}_{k \in \mathbb{N}}$ generated from (26), regardless of the convergence of the value vector sequence. This is a more general result then the classic convergence results for MDPs and robust MDPs.*

Remark 29 *Corollary 27 also implies that $\mathcal{V}^\star$ is invariant with respect to the non-stationary value iteration (26), and may prove useful in the analysis and design of MDPs with known parameter uncertainties.*

4.2 Bounds of the fixed point set

In Theorem 23, the compactness of $\mathcal{M}$ implied the compactness of $\mathcal{V}^\star$. This relationship carries over to the supremum and infimum elements of $\mathcal{M}$ and $\mathcal{V}^\star$—i.e., if $\mathcal{M}$ satisfies Assumption 9 with respect to h, then $\mathcal{V}^\star$ contains its own supremum and infimum 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,

$$\overline{V}_s := \sup_{V \in \mathcal{V}} V_s, \ \underline{V}_s := \inf_{V \in \mathcal{V}} V_s, \forall \ s \in [S]. \qquad (27)$$
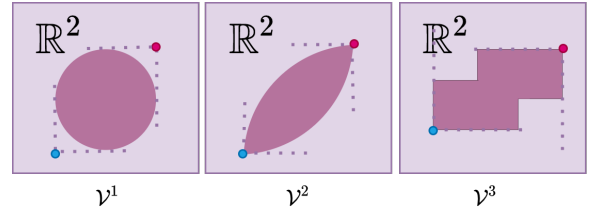

Fig. 5. The greatest least bounds of three different value function sets $\mathcal{V}^i \in \mathbb{R}^2$, where $(0,0)$ the origin is located on the lower left. Note that $\mathcal{V}^2$ and $\mathcal{V}^3$ contains their own greatest and least elements, but $\mathcal{V}^1$ does not. In $\mathcal{V}^1$, the coordinate-wise greatest and least elements are achieved by some element in $\mathcal{V}^1$ but not at the same time.

If a set $\mathcal{V} \subseteq \mathbb{R}^S$ is compact, the projection of $\mathcal{V}$ on each state s is compact. Then, the coordinate-wise supremum and infimum values for each state s are achieved by $\mathcal{V}$. However in general, no single element of the set $\mathcal{V}$ may simultaneously achieve the minimum over all states—i.e., $\overline{V}(\underline{V})$ may not be an element of $\mathcal{V}$. This is illustrated in Figure 5.

Given h and parameter uncertainty set $\mathcal{M}$, we wish to 1) bound the supremum and infimum elements of the fixed point set $\mathcal{V}^\star$ (23) and 2) derive sufficient conditions for when they are elements of $\mathcal{V}^\star$. To facilitate bounding $\mathcal{V}^\star$, we introduce the following bound operators.

Definition 30 (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*

$$\underline{h}_s(V) = \inf_{m \in \mathcal{M}} h_s(V, m), \ \overline{h}_s(V) = \sup_{m \in \mathcal{M}} h_s(V, m). \quad (28)$$

We want to bound the fixed point set $\mathcal{V}$ of the set-based value operator H (20) by the bound operators $\underline{h}/\overline{h}$ (28). First we show that $\underline{h}/\overline{h}$ are themselves α-contractive and order preserving on $\mathbb{R}^S$.
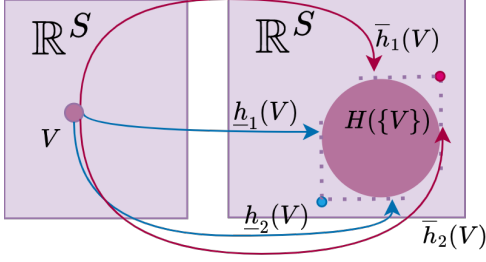
Fig. 6. We visualize the bound operator for $H(\mathcal{V})$ for a given value operator h on $\mathbb{R}^S \times \mathcal{M}$. The input set $\mathcal{V}$ is a singleton $\{V\}$ in $\mathbb{R}^2$. Here, because $\underline{h}_1$ and $\underline{h}_2$ are reached for two different parameters $m \in \mathcal{M}$, the resulting $\underline{h}(V)$ lies outside of the fixed point set.

Lemma 31 (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 $\overline{h}$ (28) are α-contractions with fixed points $\underline{X}, \overline{X}$, respectively.*

$$\overline{h}(\overline{X}) = \overline{X}, \quad \underline{h}(\underline{X}) = \underline{X}, \ \underline{X}, \overline{X} \in \mathbb{R}^S. \qquad (29)$$

PROOF. From Lemma 8, 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 $\underline{h}_s(Y) = h_s(Y, \hat{m}(s))$ and $\underline{h}_s(X) \leqslant h_s(X, \hat{m}(s))$. We upper-bound $\underline{h}_s(X) - \underline{h}_s(Y)$ by $h_s(X, \hat{m}(s)) - h_s(Y, \hat{m}(s))$, and use the α-contraction property of h to derive

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

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 $\hat{m}(s) = \sup_{m \in \mathcal{M}} h_s(X, m)$. The existence of $\underline{X}(\overline{X})$ follows from applying Banach's fixed point theorem. $\square$

Lemma 32 (Order Preservation) *The bound operators $\underline{h}$ and $\overline{h}$ (28) are order-preserving on $\mathbb{R}^S$ (Definition 3).*

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

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 $\overline{h}(\cdot) = \sup_{m \in \mathcal{M}} h(\cdot, m)$. $\square$

We show that the fixed points $\underline{X}$ and $\overline{X}$ bounds the fixed point set $\mathcal{V}^\star$ of the set-based value operator H (20).

Theorem 33 (Bounding fixed point sets) *If h (6) is a value operator on $\mathbb{R}^S \times \mathcal{M}$ and $\mathcal{M}$ is compact,*

$$\underline{X} \leqslant V \leqslant \overline{X}, \ \forall V \in \mathcal{V}^\star, \qquad (30)$$

where $\underline{X}$ and $\overline{X}$ (29) are the fixed points of the bound operators $\underline{h}$ and $\overline{h}$ (28), and $\mathcal{V}^\star$ is the fixed point set of the set-based value operator H (20) induced by h (6) on $\mathbb{R}^S \times \mathcal{M}$.

PROOF. For $\mathcal{V}^0 = \{\underline{X}, \overline{X}\}$ and $\mathcal{V}^{k+1} = H(\mathcal{V}^k)$ (25), we first show

$$\underline{X} \leqslant V \leqslant \overline{X}, \ \forall V \in \mathcal{V}^k, \qquad (31)$$

via induction. Suppose that (31) 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 (31) holds for $\mathcal{V}^{k+1}$.

Next, we show that $\underline{X}$ and $\overline{X}$ bounds the fixed point set $\mathcal{V}^\star$ for the h-induced operator H (20). From Lemma 44, 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 (30) holds. $\square$

When Assumption 9 is satisfied, the fixed point of H (20) contains its own supremum and infimum.

Theorem 34 *If h (6) on $\mathbb{R}^S \times \mathcal{M}$ satisfies Assumption 9, then there exists $\underline{m}, \overline{m} \in \mathcal{M}$ such that $\underline{h}$ and $\underline{h}$ (28) and their fixed points $\underline{X}$ and $\overline{X}$ (29) satisfies*

$$\underline{h}(\underline{X}) = h(\underline{X}, \underline{m}) = \underline{X}, \ \overline{h}(\overline{X}) = h(\overline{X}, \overline{m}) = \overline{X}. \quad (32)$$

Additionally, $\underline{X}$ and $\overline{X}$ are the least and the greatest elements of H's fixed point set $\mathcal{V}^\star$, $\underline{V}^\star$, $\overline{V}^\star$ (27) respectively, and both belong to $\mathcal{V}^\star$ (23).

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

PROOF. From Theorem 33, $\underline{X}$ 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 9, there exists $\underline{m}, \overline{m} \in \mathcal{M}$ such that $h_s(\underline{X}, \underline{m}) = \min_{m \in \mathcal{M}} h_s(\underline{X}, m)$ and $h_s(\overline{X}, \overline{m}) = \min_{m \in \mathcal{M}} h_s(\overline{X}, m)$ 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 26 to conclude that $\underline{X}, \overline{X} \in \mathcal{V}^\star$. $\square$

5 Revisiting robust MDP

We re-examine robust MDP with the set-theoretical analysis in this section, and show that Assumption 9 generalizes the rectangularity assumption made in robust MDPs, thus enabling robust dynamic programming techniques to be available to a wider class of MDP problems and contraction operators.

Recall the optimistic value vector $W^o \in \mathbb{R}^S$ and robust value vectors $W^r \in \mathbb{R}^S$ of a discounted MDP $([S], [A], C, P, \gamma)$ from [8, 15] as the fixed points of the following operators.

$$W_s^o = \min_{\pi_s \in \Delta_A} \min_{(C,P) \in \mathcal{M}} g_s^\pi(W^o, C, P), \forall s \in [S] \quad (33)$$

$$W_s^r = \min_{\pi_s \in \Delta_A} \max_{(C,P) \in \mathcal{M}} g^\pi(W^r, C, P), \forall s \in [S] \quad (34)$$

The optimistic policy π^o and robust policy π^r are the optimal policies corresponding to (33) and (34), respectively.

$$\pi_s^o \in \operatorname*{argmin}_{\pi_s \in \Delta_A} \min_{(C,P) \in \mathcal{M}} g_s^\pi(W^o, C, P), \forall s \in [S] \quad (35)$$

$$\pi_s^r \in \operatorname*{argmin}_{\pi_s \in \Delta_A} \max_{(C,P) \in \mathcal{M}} g_s^\pi(W^r, C, P), \forall s \in [S] \quad (36)$$

For readability, we denote the policy evaluation operator under π^o as g^o and the policy evaluation operator under π^r as g^r.

When $\mathcal{M}$ is (s, a)-rectangular (16), the set of policies satisfying (35) and (36) are non-empty and includes deterministic policies [8, Thm 3.1]. When $\mathcal{M}$ is s-rectangular and convex, the set of policies satisfying (36) is non-empty but may be mixed [24, Thm 4]. When $\mathcal{M}$ is convex, we show that policies (35) and (36) exist.

Proposition 35 *If the MDP parameter set $\mathcal{M}$ is compact and convex, then*

(1) W^o (33) and W^r (34) exist and satisfy $\overline{f}(W^r) = W^r$, $\underline{f}(W^o) = W^o$, where $\overline{f}$ and $\underline{f}$ (28) are the bound operators of the Bellman operator (10).
(2) π^o (35) and π^r (36) exist.

PROOF. Recall the Bellman operator f (6). When $\mathcal{M} \times \Delta_A$ is compact, the formulation of the fixed point of $\underline{f}$ (28) is equivalently given by

$$\underline{f}(\underline{X}) = \min_{(C,P) \in \mathcal{M}} \min_{\pi_s \in \Delta_A} g_s^\pi(\underline{X}, C, P), \ \forall s \in [S]. \quad (37)$$

We note that (37) is identical to the formulation of W^o (33). Therefore, $W^o = \underline{X}$ is the fixed point of $\underline{f}$. When $\mathcal{M}$ is compact, W^o exists due to Lemma 31. From (35), 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 $\pi^o \in \Pi$ exists.

For the robust scenario: when $\mathcal{M}$ is compact, the fixed point of $\overline{f}$ (28), $\overline{X}$, exists from Lemma 31 and is given by

$$\overline{X}_s = \max_{(C,P) \in \mathcal{M}} \min_{\pi_s \in \Delta_A} g_s^\pi(\overline{X}, C, P), \ \forall s \in [S]. \quad (38)$$

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 [14] to switch the order of min and max in (38) to derive

$$\overline{X}_s = \min_{\pi_s \in \Delta_A} \max_{(C,P) \in \mathcal{M}} g_s^\pi(\overline{X}, C, P), \ \forall s \in [S]. \quad (39)$$

Equation (39) is identical to (34), therefore $W^r = \overline{X}$ and exists by Lemma 31. In (39), $\max_{(C,P) \in \mathcal{M}} g_s^\pi(\overline{X}, C, P)$ 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. $\square$

Remark 36 *Since $\max_{(C,P) \in \mathcal{M}} g_s^\pi(\overline{X}, C, P)$ is piece-wise linear in π_s, the optimal π_s^r is mixed policy in general. This is consistent with the results in [24].*

Proposition 35 generalizes the results from [24] to show that (34) exists when $\mathcal{M}$ is compact and convex instead of s-rectangular and convex. From Theorem 33, W^o and W^r bound of the fixed point sets of the π^o and π^r. They become infimum and supremum elements when $\mathcal{M}$ satisfies Assumption 9 with respect to g^o and g^r. We explicitly derive this result next. First, we introduce some notations: let $G^o = G^{\pi_o}$, the fixed point of G^o be $\mathcal{V}^o$, $G^r = G^{\pi^r}$, and the fixed point of G^r be $\mathcal{V}^r$.

$$\mathcal{V}^o = \{g^o(V, C, P) \mid (C, P) \in \mathcal{M}, V \in \mathcal{V}^o\}, \quad (40)$$

$$\mathcal{V}^r = \{g^r(V, C, P) \mid (C, P) \in \mathcal{M}, V \in \mathcal{V}^r\}. \quad (41)$$

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

$$\underline{V}_s^r = \min_{V \in \mathcal{V}^r} V_s, \ \overline{V}_s^r = \max_{V \in \mathcal{V}_r} V_s, \ \forall s \in [S]. \quad (42)$$

$$\underline{V}_s^o = \min_{V \in \mathcal{V}^o} V_s, \ \overline{V}_s^o = \max_{V \in \mathcal{V}^o} V_s, \ \forall s \in [S]. \qquad (43)$$

We compare these with the fixed point set of the Bellman operator, $\mathcal{V}^B = \{\min_\pi g^\pi(V, C, P) \mid (C, P) \in \mathcal{M}, V \in \mathcal{V}^B\}$ (23), denoted by $\overline{V}^B$ and $\underline{V}^B$ as

$$\underline{V}_s^B = \min_{V \in \mathcal{V}^B}[V]_s, \ \overline{V}_s^B = \max_{V \in \mathcal{V}^B} V_s, \ \forall s \in [S]. \qquad (44)$$

Our next result proves the relationship between $\underline{V}^B, \underline{V}^o, \underline{V}^r, \overline{V}^B, \overline{V}^o, \overline{V}^r$ when f, g^o, and g^r on $\mathbb{R}^S \times \mathcal{M}$ satisfy Assumption 9.

Theorem 37 *If f, g^o, g^r satisfy Assumption 9 on $\mathbb{R}^S \times \mathcal{M}$, then the bounding elements (43) (42) (44) of the corresponding fixed point sets $\mathcal{V}^B, \mathcal{V}^o$ (40) and $\mathcal{V}^r$ (41) are ordered as*

$$\underline{V}^B = \underline{V}^o \leqslant \underline{V}^r, \ \overline{V}^B = \overline{V}^r \leqslant \overline{V}^o. \qquad (45)$$

PROOF. Since $\underline{V}^o$ is the infimum element for the fixed point set $\mathcal{V}^o$ (43), we can apply Theorem 34 to derive

$$\underline{V}^o = \min_{(C,P) \in \mathcal{M}} g^o(\underline{V}^o, C, P). \qquad (46)$$

By definition of π_o (35), $\min_{(C,P) \in \mathcal{M}} g^o(\underline{V}^o, C, P) = \min_{(C,P) \in \mathcal{M}} \min_{\pi \in \Pi} g^\pi(\underline{V}^o, C, P)$. As the two minima commute,

$$\min_{(C,P) \in \mathcal{M}} g^o(\underline{V}^o, C, P) = \min_{(C,P) \in \mathcal{M}} \min_{\pi \in \Pi} g^\pi(\underline{V}^o, C, P). \qquad (47)$$

Combining (46) and (47), $\underline{V}^o$ is exactly the unique fixed point of $\min_{(C,P) \in \mathcal{M}} \min_{\pi \in \Pi} g^\pi(\cdot, C, P)$. However, by applying Theorem 34 to f on $\mathbb{R}^S \times \mathcal{M}$, $\underline{V}^B$ is also the unique fixed point of $\min_{(C,P) \in \mathcal{M}} \min_{\pi \in \Pi} g^\pi(\cdot, C, P)$. Therefore $\underline{V}^o = \underline{V}^B$.

From (42), $\underline{V}^r = \min_{(C,P) \in \mathcal{M}} g^r(\underline{V}^r, C, P)$, we can minimize over the policy space to lower bound $\underline{V}^r$ as

$$\underline{V}^r \geqslant \min_{\pi \in \Pi} \min_{(C,P) \in \mathcal{M}} g^r(\underline{V}^r, C, P). \qquad (48)$$

Since the right hand side of (48) is equivalent to $f(\underline{V}^r)$, (48) is equivalent to $\underline{V}^r \geqslant f(\underline{V}^r)$. From Lemma 32, f is order-preserving in V, we conclude that $\underline{V}^o = \underline{V}^\star \leqslant \underline{V}^r$.

From Theorem 34, $\overline{V}^r$ is the fixed point of $\overline{g^r}$, such that

$$\overline{V}^r = \max_{(C,P) \in \mathcal{M}} g^r(\overline{V}^r, C, P). \qquad (49)$$

We apply $\min_\pi$ to both sides of (49) and use the definition of π_r to derive that $\overline{V}^r$ is the fixed point of

$\min_{\pi \in \Pi} \max_{(C,P) \in \mathcal{M}} g^\pi(V^r, C, P)$. From Assumption 9, there exists $(\overline{C}, \overline{P}) \in \mathcal{M}$ that maximizes $g^\pi(\overline{V}, C, P)$, so $\overline{V}^r$ equivalently satisfies

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

From Corollary 26, this implies that $\overline{V}^r \in \mathcal{V}^B$ and therefore $\overline{V}^r \leqslant \overline{V}^B$. Next we show $\overline{V}^B \leqslant \overline{V}^r$. From Theorem 34, $\overline{V}^B$ 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,

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

Since $\pi_r \in \Pi$,

$$\overline{V}^B \leqslant \max_{(C,P) \in \mathcal{M}} g^r(\overline{V}^B, C, P). \qquad (50)$$

The right-hand side of (50) is $\overline{g}^r(\overline{V}^B)$ (28), such that (50) is equivalent to $\overline{V}^B \leqslant \overline{g}^r(\overline{V}^B)$. We 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 32, $\overline{g}^r$ is order preserving. Therefore $\overline{V}^B = V^1 \leqslant V^r$.

Finally, Theorem 34 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, $\overline{V}^o \geqslant \min_{\pi \in \Pi} \max_{(C,P) \in \mathcal{M}} g^\pi(\overline{V}^o, C, P)$. From the min-max inequality,

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

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 [17, Thm.6.2.2], we conclude that $\overline{V}^o \geqslant \overline{V}^B$. $\square$

Remark 38 *Through our fixed-point analysis, we see 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 $\mathcal{M}$.*

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 9.

Corollary 39 (Robust MDP under Assumption 9) *If $\mathcal{M}$ is compact and convex, and f, g^o, g^r satisfy Assumption 9 on $\mathbb{R}^S \times \mathcal{M}$, then W^o (33) and W^r (34) are the infimum and supremum value vectors for the*
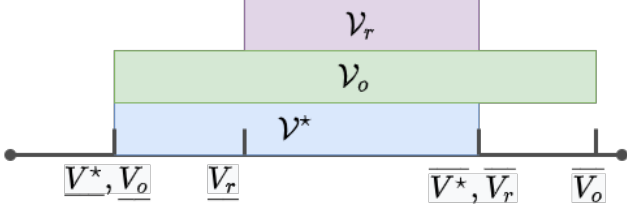
Fig. 7. Illustration of Theorem 37. The purple, green, blue regions indicate the ranges of $\mathcal{V}^r$, $\mathcal{V}^o$, and $\mathcal{V}^B$, respectively.

policy evaluation operator under π^o (35) and π^r (36), respectively.

$$W_s^o = \inf_{V \in \mathcal{V}^o} [V]_s, W_s^r = \sup_{V \in \mathcal{V}^r} [V]_s, \ \forall s \in [S], \qquad (51)$$

where $\mathcal{V}^o$ (40) and $\mathcal{V}^r$ (41) are the fixed point sets of policies π^o and π^r under parameter uncertainty $\mathcal{M}$, respectively.

PROOF. When f satisfies Assumption 9 on $\mathbb{R}^S \times \mathcal{M}$, Theorem 34 shows that $\underline{V}^B = W^o$, $\overline{V}^B = W^r$. If f, g^o, and g^r also satisfies Assumption 9 on $\mathbb{R}^S \times \mathcal{M}$, then we apply Theorem 37 to derive $W^o = \underline{V}^o$ and $W^r = \overline{V}^r$. This proves the corollary statement. $\square$

Remark 40 *When Assumption 9 is not satisfied, W^o and W^r still bound $\underline{V}^o$ and $\overline{V}^r$. This result is also stated in [24].*

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 through our techniques. 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 (25), 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 equivalent to the one-step output of the bound operators $\underline{h}$ and $\overline{h}$ (28) applied to the extremal points of $\mathcal{V}$.

Theorem 41 (One step H bounds) *Consider a set operator H (20) and its bound operators $\underline{h}$ and $\overline{h}$ (28) 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 $\overline{h}(\overline{V})$ (28) as*

$$\underline{h}(\underline{V}) \leqslant V \leqslant \overline{h}(\overline{V}), \quad \forall \ V \in H(\mathcal{V}). \qquad (52)$$

where $\underline{V}$ and $\overline{V}$ (27) are the extremal elements of $\mathcal{V}$. If h satisfies Assumption 9 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

$$\underline{h}_s(\underline{V}) = \inf_{(V,m) \in \mathcal{V} \times \mathcal{M}} h_s(V, m), \ \overline{h}_s(\overline{V}) = \sup_{(V,m) \in \mathcal{V} \times \mathcal{M}} h_s(V, m).$$
$$(53)$$

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 32) such that $\overline{h}_s(V) \leqslant \overline{h}_s(\overline{V})$ for all $V \in \mathcal{V}$. We conclude that

$$h(V, m) \leqslant \overline{h}(\overline{V}), \ \forall (V, m) \in \mathcal{V} \times \mathcal{M}. \qquad (54)$$

Since $\overline{h}$ is an upper bound, and sup is the least upper bound, it holds that $\sup_{V,m}[h(V, m)]_s \leqslant \overline{h}(\overline{V})$. We use the definition of $H(\mathcal{V})$ (20) 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 9 on $\mathbb{R}^S \times \mathcal{M}$ and $\underline{V}, \overline{V} \in \mathcal{V}$, Assumption 9 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. $\square$

Based on Theorem 41, we propose the following bound approximation algorithm of the fixed point set $\mathcal{V}^\star$ (23) for a set-valued operator H (6).

Algorithm 1 Bound approximation of the fixed point set $\mathcal{V}$

Input: $\mathcal{C}, \mathcal{P}, V^0, \epsilon$.
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: $\quad e^{k+1} = \max \left\{ \left\| \underline{V}^{k+1} - \underline{V}^k \right\|, \left\| \overline{V}^{k+1} - \overline{V}^k \right\| \right\}$
7: $\quad k = k + 1$
8: **end while**

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 we discuss solution methods for different shapes of $\mathcal{M}$.

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

$$\underline{V}^{k+1} = \min \left\{ h_s(\underline{V}^k, m_i) \mid i = \{1, \ldots, N\} \right\}. \quad (55)$$

For line 5, we replace min with max in (55).

(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 [24]. When $\mathcal{M}$ can be characterized by an ellipsoidal set of parameters, the solutions to lines 4 and 5 is given in [24].

We recall the stochastic path planning problem from Example 17 with the two different parameter uncertainty scenarios. When the wind field uncertainty is discrete, $\mathcal{M}$ is finite, when wind field is a combination of the major wind trends, $\mathcal{M}$ is convex.

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 9 with respect to h, Algorithm 1 derives the exact bounds of $\mathcal{V}$. Algorithm 1 has similar rates of convergence in Hausdorff distance as standard value iteration using h on $\mathbb{R}^S$.

Theorem 42 *Consider the value operator h, compact uncertainty set $\mathcal{M}$, and the fixed point set $\mathcal{V}^\star$ of the set-based operator H (20) induced by h on $\mathbb{R}^S \times \mathcal{M}$. If $\mathcal{M}$ satisfies Assumption 9 with respect to h, then at each iteration k,*

$$\left\| \underline{V}^{k+1} - \underline{V}^\star \right\| \leqslant \alpha \left\| \underline{V}^k - \underline{V}^\star \right\|, \left\| \overline{V}^{k+1} - \overline{V} \right\| \leqslant \alpha \left\| \overline{V}^k - \overline{V}^\star \right\| \quad (56)$$

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

$$\max\{ \left\| \underline{V}^k - \underline{V}^\star \right\|, \left\| \overline{V}^k - \overline{V}^\star \right\| \} < \epsilon. \quad (57)$$

PROOF. From Algorithm 1, $\overline{V}^{k+1} = \overline{h}(\overline{V}^k)$. From Lemma 31, $\overline{h}$ is an α-contraction. We obtain $\left\| \overline{V}^{k+1} - \overline{V}^\star \right\| \leqslant \alpha \left\| \overline{V}^k - \overline{V}^\star \right\|$ and note that (56) holds by induction. Next, we apply triangle inequality to $\left\| \overline{V}^k - \overline{V}^\star \right\|$ to derive

$$\left\| \overline{V}^k - \overline{V}^\star \right\| \leqslant \left\| \overline{V}^k - \overline{V}^{k+1} \right\| + \left\| \overline{V}^{k+1} - \overline{V}^\star \right\|. \quad (58)$$

We can then use $\left\| \overline{V}^{k+1} - \overline{V}^\star \right\| \leqslant \alpha \left\| \overline{V}^k - \overline{V}^\star \right\|$ to bound (58) as $\left\| \overline{V}^k - \overline{V}^\star \right\| \leqslant \frac{1}{1-\alpha} \left\| \overline{V}^k - \overline{V}^{k+1} \right\|$. A similar argument can show that $\left\| \underline{V}^k - \underline{V}^\star \right\| \leqslant \frac{1}{1-\alpha} \left\| \underline{V}^k - \underline{V}^{k+1} \right\|$. When Algorithm 1's while condition is satisfied, $\max \left\{ \left\| \overline{V}^k - \overline{V}^\star \right\|, \left\| \underline{V}^k - \underline{V}^\star \right\| \right\} \leqslant \epsilon$. This concludes our proof. $\square$

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 41 applies with $\alpha = \gamma$.

Remark 43 *Theorem 42 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 Time-varying Wind Fields

We apply set-based value iteration to wind-assisted probabilistic path planning of a balloon in strong, uncertain wind fields [25]. MDP as a model for wind-assisted path planning of balloons in the stratosphere and exoplanets has recently gained traction [2, 25]. Discrete state-action MDPs have been shown to be a viable high-level path planning model [25] for such applications.

Mission Objective. In the two dimensional wind-field, we assume that the wind-assisted balloon is tasked with reaching target state $(8, 8)$ in Figure 8 using minimum fuel.

Uncertain Wind Fields. By collecting a set of 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 the framework described in [25], we model the path planning problem in an uncertain wind field as an infinite horizon, discounted MDP with discrete state-actions in a two-dimensional space. 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. As a result, we can disregard the vertical position during planning.
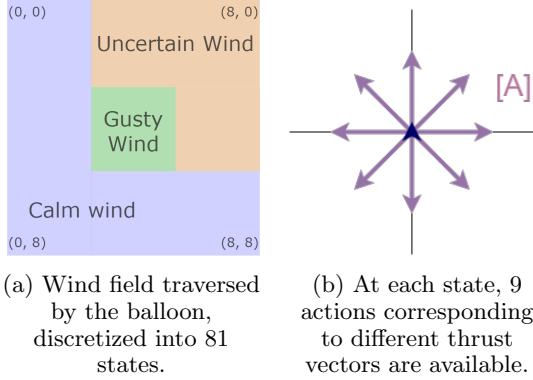

(a) Wind field traversed by the balloon, discretized into 81 states.

(b) At each state, 9 actions corresponding to different thrust vectors are available.

Fig. 8.

States. A total of 81 states represent the two-dimensional space, composed of three different regions characterized by their wind variability as shown in Figure 8.

(1) **Calm wind**. In calm states S_{calm}, the wind magnitude varies uniformly between $[0, 0.5]$, and the wind direction is uniformly sampled between $[0, 2\pi]$. $S_{calm} = \{(i, j) \mid (0, 0) \leqslant (i, j) \leqslant (2, 8), (6, 0) \leqslant (i, j) \leqslant (8, 8)\}$.

(2) **Gusty wind**. In states with gusts S_{gusty}, wind magnitude is consistently 1, while the wind direction is uniformly sampled between $[0, 2\pi]$. $S_{gusty} = \{(i, j) \mid (3, 3) \leqslant (i, j) < (6, 6)\}$.

(3) **Unreliable wind**. In unreliable states $S_{unreliable}$, a predictable wind front occasionally moves across an otherwise windless region. In other words, the wind magnitude is either 0 or 1 and the wind direction varies uniformly between $[\pi/4, \pi/2]$.

Actions. The balloon is equipped with an actuator that provides a constant thrust of 1 in 8 discretized directions shown in Figure 8b. The the only stationary action vector with zero magnitude is highlighted in blue in the center of Figure 8b. We assume that the 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 wind effects.

Transition Probabilities. The transition probabilities in S_{calm} and S_{gusty} are certain. At each state s, we consider the following neighboring states.

(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 adjacent states as shown in Figure 9a.

(a) State transition in calm wind.

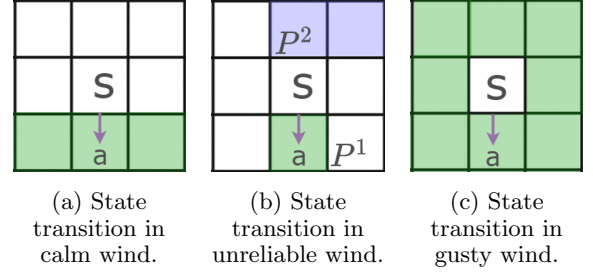(b) State transition in unreliable wind.

(c) State transition in gusty wind.

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

(4) $\mathcal{N}(s, a, 2)$: the up and upper-right neighbors of s, as shown in Figure 9b.

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

$$P_{sa,s'} = \begin{cases} \frac{1}{\mathcal{N}(s,a,1)}, & s' \in \mathcal{N}(s, a, 1) \\ 0 & \text{otherwise} \end{cases}, \ \forall \ s \in [S_{calm}].$$
(59)

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

$$P_{sa,s'} = \begin{cases} \frac{1}{\mathcal{N}(s)}, & s' \in \mathcal{N}(s) \\ 0 & \text{otherwise} \end{cases}, \ \forall \ s \in [S_{gusty}], \ \forall \ a \in [A].$$
(60)

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

$$P_{sa,s'}^1 = \begin{cases} 1, & s' \in \mathcal{N}(s, a, 0) \\ 0 & \text{otherwise} \end{cases}, \ \forall \ s \in [S_{gusty}], \ \forall \ a \in [A].$$
(61)

$$P_{sa,s'}^2 = \begin{cases} 0.5, & s' \in \mathcal{N}(s, a, 2) \\ 0 & \text{otherwise} \end{cases}, \ \forall s \in [S_{gusty}], \ \forall a \in [A].$$
(62)

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

$$\mathcal{P}_s = \{P_{sa}^i \mid i \in \{1, 2\}, a \in [A]\}, \ \forall s \in [S_{unreliable}]. \quad (63)$$

Cost. We define the following state-action cost to achieve the mission objective: 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 given action.

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

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

7.1 Bellman, optimistic policy, and robust policy

We first compute the optimistic and robust bounds of the MDP with parameter uncertainty in $\mathcal{P}$ when $s \in [S_{unreliable}]$ by running Algorithm 1. The results are shown in Figure 10.
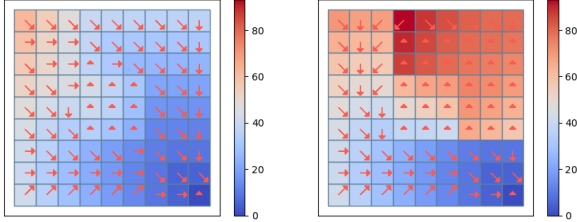
(a) Optimistic case with expected objective of 54.2

(b) Robust case with expected objective of 96.7.

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 $\mathcal{V}^o$ (40) and $\mathcal{V}^r$ (41) 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.

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

Table 1
Bellman, optimistic policy, robust policy value bounds of the uncertain wind field.

Time-varying wind field Next, we consider a time-varying wind field: at each time step k, the transition probability P^k is chosen at random from $\mathcal{P}$ (63). In this time-varying wind field, we compare three different policy deployments: 1) stationary optimistic policy π^o as policy operator g^o (40), 2) stationary robust policy π^r as policy operator g^r (41), and 3) dynamically changing policy that is optimal for the MDP $([S],[A],P^k,C,\gamma)$ as f (10). These three different policy deployments are given by

$$V^{k+1} = g^o(V^k, C, P^k), \tag{64}$$

$$V^{k+1} = g^r(V^k, C, P^k), \tag{65}$$

$$V^{k+1} = f(V^k, C, P^k). \tag{66}$$

The resulting cost-to-go at state $s_{orig} = [0,0]$ is plotted in Figure 11. Here, we see that the optimistic policy deployment (64) has the greatest variation in value over the course of 50 MDP time steps. Both the robust policy deployment (65) and the dynamically changing policy deployment (66) achieve better upper-bound at

each MDP iteration. The dynamically changing policy deployment (66) achieves less than 70 in cost-to-go on average, which is the best among all three deployments. As we discussed in Remark 38, the robust policy deployment 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 Figure 12.

8 Conclusion

In this paper, we categorized a class of operators utilized to solve Markov decision processes as value operators and lifted their input space from vectors to *compact sets* of vectors. We showed using fixed point analysis that the set extensions of value operators have fixed point sets which remain invariant given a compact set of MDP parameter uncertainties. These sets were applied to robust dynamic programming to further enrich existing results and generalize the k-rectangularity assumption for robust MDPs. Finally, we applied our results to a path planning problem for time-varying wind fields. For future work, we plan on applying set-based value operators to stochastic games in the presence of uncoordinated players such as humans, as well as applying value operators to reinforcement learning to synthesize robust learning algorithms.

References

[1] Wesam H Al-Sabban, Luis F Gonzalez, and Ryan N Smith. Wind-energy based path planning for unmanned aerial vehicles using markov decision processes. In *2013 IEEE International Conference on Robotics and Automation*, pages 784–789. IEEE, 2013.

[2] Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

[3] Marc G Bellemare, Georg Ostrovski, Arthur Guez, Philip Thomas, and Rémi Munos. Increasing the action gap: New operators for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

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

[5] Robert Givan, Sonia Leach, and Thomas Dean. Bounded-parameter markov decision processes. *Artificial Intelligence*, 122(1-2):71–109, 2000.
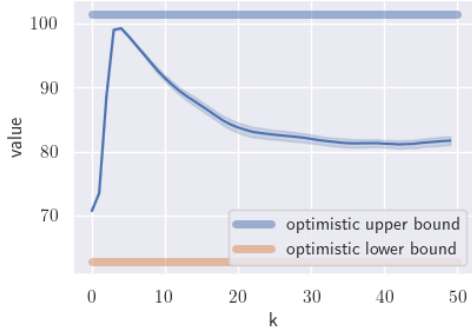
(a) Optimistic Policy with $\mathcal{V}^o$'s bounding values.
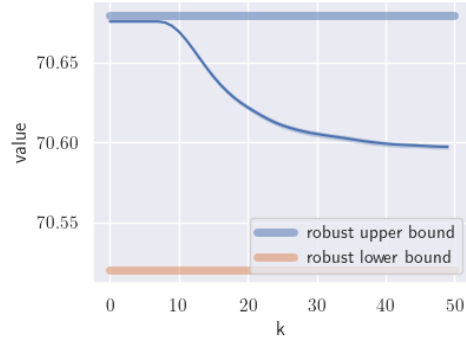
(b) Robust Policy with $\mathcal{V}^r$'s bounding values.
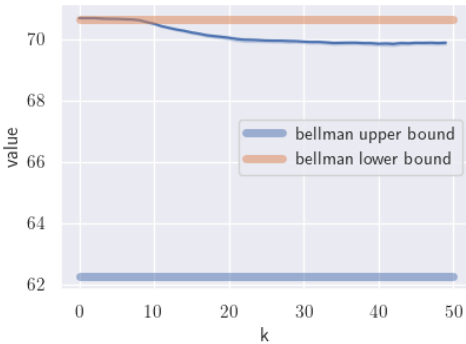
(c) Dynamically changing policy with $\mathcal{V}^B$'s bounding values.

Fig. 11. Comparison of robust policy, optimistic policy, and Bellman policy's value trajectories in time-varying 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 supremum and infimum values of the fixed points.
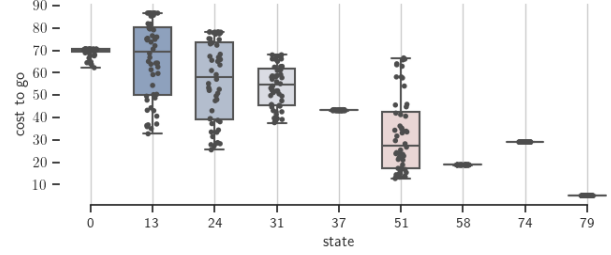

Fig. 12. Comparison of different optimal value vectors under the Bellman operator for 50 randomly sampled MDPs. On the x-axis, the state number is computed as $i \times 9 + j$.

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

[9] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

[10] Erwan Lecarpentier and Emmanuel Rachelson. Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

[11] Sarah HQ Li, Assalé Adjé, Pierre-Loïc Garoche, and Behçet Açıkmeşe. Bounding fixed points of set-based bellman operator and nash equilibria of stochastic games. *arXiv preprint arXiv:2001.07889*, 2020.

[12] Shie Mannor, Ofir Mebel, and Huan Xu. Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

[13] Francisco S Melo. Convergence of q-learning: A simple proof. *Institute Of Systems and Robotics, Tech. Rep*, pages 1–4, 2001.

[14] J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

[15] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[16] Sindhu Padakandla, Prabuchandran KJ, and Shalabh Bhatnagar. Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, 50(11):3590–3606, 2020.

[17] Martin L Puterman. *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[18] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[19] Bernd SW Schröder. Ordered sets. *Springer*, 29:30, 2003.

[20] Csaba Szepesvári and Michael L Littman. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural computation*, 11(8):2017–2060, 1999.

[21] Gerald Tesauro et al. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[22] Herke Van Hoof, Tucker Hermans, Gerhard Neumann, and Jan Peters. Learning robot in-hand manipulation with tactile features. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 121–127. IEEE, 2015.

[23] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

[6] Vineet Goyal and Julien Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

[7] Jeff Henrikson. Completeness and total boundedness of the hausdorff metric. In *MIT Undergraduate Journal of Mathematics*. Citeseer, 1999.

[24] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[25] Michael T Wolf, Lars Blackmore, Yoshiaki Kuwata, Nanaz Fathpour, Alberto Elfes, and Claire Newman. Probabilistic motion planning of balloons in strong, uncertain wind fields. In *2010 IEEE International Conference on Robotics and Automation*, pages 1123–1129. IEEE, 2010.

[26] Insoon Yang. A convex optimization approach to distributionally robust markov decision processes with wasserstein distance. *IEEE control systems letters*, 1(1):164–169, 2017.

Lemma 44 *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 $\|\cdot\|$.*

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