Automatica 130 (2021) 109685

Contents lists available at ScienceDirect

Automatica

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

Bounding fixed points of set-based Bellman operator and Nash equilibria of stochastic games ☆

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

[a] William E.Boeing Department of Aeronautics and Astronautics, University of Washington, Seattle, USA
[b] LAMPS, Université de Perpignan Via Domitia, Perpignan, France
[c] ENAC, Université de Toulouse, Toulouse, France

ARTICLE INFO

Article history:
Received 21 January 2020
Received in revised form 6 February 2021
Accepted 7 April 2021
Available online 13 May 2021

Keywords:
Markov decision process
Learning theory
Stochastic control
Multi-agent systems
Learning in games
Decision making and autonomy

ABSTRACT

Motivated by uncertain parameters encountered in Markov decision processes (MDPs) and stochastic games, we study the effect of parameter-uncertainty on Bellman operator-based algorithms under a set-based framework. Specifically, we first consider a family of MDPs where the cost parameters are in a given compact set; we then define a Bellman operator acting on a set of value functions to produce a new set of value functions as the output under all possible variations in the cost parameter. We prove the existence of a *fixed point* of this set-based Bellman operator by showing that the operator is contractive on a complete metric space, and explore its relationship with the corresponding family of MDPs and stochastic games. Additionally, we show that given interval set-bounded cost parameters, we can form exact bounds on the set of optimal value functions. Finally, we utilize our results to bound the value function trajectory of a player in a stochastic game.

1. Introduction

Markov decision process (MDP) is a fundamental framework for control design in stochastic environments, reinforcement learning, and stochastic games (Açıkmeşe & Bayard, 2015; Demir, Eren, & Açikmeşe, 2015; Filar & Vrieze, 2012; Li, Yu, Calderone, Ratliff, & Acikmese, 2019). With known cost and transition probabilities, solving an MDP is equivalent to minimizing an objective in expectation, and requires determining the optimal value function as well as deriving the corresponding optimal policy for each state. Relying on the fact that the optimal value function is the *fixed point* of the Bellman operator, dynamic programming methods iteratively apply variants of the Bellman operator to converge to the optimal value function and the optimal policy (Puterman, 2014).

We are motivated to study MDPs where the parameters that define the environment are *sets* rather than single-valued. Such a set-based perspective arises naturally in the analysis of parameter-uncertain MDPs and stochastic games. In this paper, we develop a framework for evaluating MDPs on compact sets of costs and value functions. Specifically, we show that when the cost parameter of the MDP is in a compact set rather than being single-valued, we can define a Bellman operator that is contractive with respect to the Hausdorff distance on the space of compact sets. We prove the existence of a unique and compact *fixed point set* that the operator must converge to, and give interpretations of the fixed point set in the context of parameter-uncertain MDPs and stochastic games.

When modelling a system as a stochastic process, sampling techniques are often used to determine cost and transition probability parameters. In such scenarios, the MDP can be either interpreted as a standard MDP with error bounds on its parameters or as a *set-based MDP* in which its parameters are set-based rather than single-valued. In the former approach, the MDP can be solved with standard dynamic programming methods, and the stability of its solution with respect to parameter perturbation can be analysed locally (Abbad & Filar, 1992; Altman & Gaitsgory, 1993; Bielecki & Filar, 1991). However, these sensitivity results are only local approximations in the context of compact parameter sets. The latter approach is not well explored — some research exists on bounded interval set MDPs (Givan, Leach, & Dean, 2000), in which dynamic programming techniques such as value and policy iteration have been shown to converge. However, while it is known that parameter-uncertain MDPs may result in value function sets such as polytopes (Dadashi, Bellemare, Taïga, Roux, & Schuurmans, 2019), there are no convergence guarantees for

☆ This work has been partially supported by Feanicses project, France ANR-17-CE25-0018 and NSF, USACNS-1736582. The material in this paper was presented at the 21st IFAC World Congress (IFAC 2020), July 12–17, 2020, Berlin, Germany. This paper was recommended for publication in revised form by Associate Editor Alessandro Abate under the direction of Editor Ian R. Petersen.

* Corresponding author.
E-mail addresses: sarahli@uw.edu (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.2021.109685
0005-1098/

dynamic programming with polytopic sets of value functions. In this paper, we show that for a set-based MDP with a compact set of cost parameters, and any regular MDP whose cost is an element of the said compact set of cost parameters, the associated set-based Bellman operator has a unique and compact fixed point set that must contain the optimal value function of the regular MDP.

As opposed to parameter-uncertain MDPs where the underlying cost and probability parameters are constant albeit uncertain, stochastic games generate MDPs where the cost and probability parameters *vary* with opponents' changing policies. An individual player can interpret a stochastic game as an MDP with a parameter-varying environment, where holding all opponents' policies fixed, the stochastic game played by player i is equivalent to a regular MDP. At a fixed joint policy, we say that a player's policy is *optimal* if it is optimal with respect to the corresponding MDP. If every player's policy is optimal with respect to their opponents' fixed policies within a joint policy, then we say that the game has reached a Nash equilibrium — i.e., every player's policy is *optimal* for the current joint policy. A Nash equilibrium defines a joint policy at which no player has any incentive to unilaterally deviate from. In learning theory for stochastic games, it is often each player's goal to achieve the Nash equilibrium through an iterative process. Therefore, many learning algorithms are based on variants of dynamic programming, where each player solves an MDP with costs and transition probabilities changing at each iteration (Bu, Babu, De Schutter, et al., 2008; Littman, 2001). In this paper, we apply a set-based dynamic programming technique to a single-controller stochastic game. However, rather than demonstrating convergence to a Nash equilibrium, we show that the *set* of Nash equilibria must be contained in the fixed point set of a set-based Bellman operator.

In Li, Adjé, Garoche, and Açıkmeşe (2020), we began our analysis of set-based MDPs by proving the existence of a unique fixed point set associated with the set-based Bellman operator. In this paper, we demonstrate the significance of this fixed point set by relating it to the fixed points of parameter-uncertain MDPs and the Nash equilibria set of stochastic games. We further explore the fixed point set in the context of iterative solutions to stochastic games, and show that the fixed point set of the set-based Bellman operator bounds the asymptotic behaviour of dynamic programming-based learning algorithms.

The paper is structured as follows: we provide references to existing research in Section 2; we recall definitions of the MDP and the Bellman operator in Section 3; Section 4 extends these definitions to set-based MDPs, providing theoretical results for the existence of a fixed point set of a set-based Bellman operator. Section 5 relates properties of the fixed point set to stochastic games. An interval set-based MDP is presented in Section 6 with a computation of exact bounds, while the application to stochastic games is illustrated in Section 7, where we model unknown policies of the opponent as cost intervals.

2. Related research

Bounding the fixed point of the Bellman operator with uncertain parameters is well studied under robust MDPs such as in Delage and Mannor (2010) and Wiesemann, Kuhn, and Rustem (2013), where the MDP parameters are either assumed or estimated as random variables from a known Gaussian distribution (Delage & Mannor, 2010; Wiesemann et al., 2013). In contrast, we model our MDP cost parameter-uncertainty as a compact set without any probabilistic prior assumptions. Therefore, our results are absolute as opposed to chance-constrained or stochastic.

A closely related work from robust MDP is Iyengar (2005), where the author analyzes what we consider as the lower bound on the value function of parameter-uncertain MDPs and connects parameter-uncertain MDPs to stochastic games. We generalize these results for cost uncertainty only, and show that there exists an invariant set corresponding to parameter-uncertain MDPs, and that dynamic programming-based algorithms can converge to the invariant set itself instead of just obtaining a lower bound.

Our parameter-uncertain MDP model also generalizes the bounded parameter model presented in Givan et al. (2000), which considers interval sets instead of general compact sets.

Other approaches to bound the value functions of cost-uncertain MDPs include Dick, Gyorgy, and Szepesvari (2014), Haddad and Monmege (2018). MDPs with reachability objectives are studied in Haddad and Monmege (2018) under a graph-theoretical MDP uncertainty model. However, the techniques utilized in Haddad and Monmege (2018) require abstraction of the MDP state space and therefore do not directly extend to value functions that are defined per state. A learning approach to solve cost-uncertain MDPs is taken in Dick et al. (2014), and convergence in terms of regret is shown using gradient-based algorithms instead of dynamic programming approaches.

Introduced in Shapley (1953), stochastic games generalize MDPs to the multi-agent setting, where the goal of each player is to minimize their individual cost functions. Typically, this leads to stable behaviour at a Nash equilibrium. In general, it is difficult to find the Nash equilibrium of a general-sum stochastic game; the computation complexity has been shown to be NP-hard in Chatterjee, Majumdar, and Jurdziński (2004) and value iteration for such games is shown to diverge in Kearns, Mansour, and Singh (2000). Convergence guarantees for Bellman operator-based algorithms exist under limited settings such as two-player stochastic games or zero-sum stochastic games (Eisentraut, Křetínskỳ, & Rotar, 2019; Prasad, Prashanth, & Bhatnagar, 2015; Shapley, 1953; Wei, Hong, & Lu, 2017). However, the same algorithms have been shown to converge empirically in a wide range of applications including poker and cyber-security (Ganzfried & Sandholm, 2009; Shiva, Roy, & Dasgupta, 2010). In this paper, we consider single-controller stochastic games with *imperfect information* (Filar & Vrieze, 2012, Def. 6.3.6), and show that our set-based value iteration algorithm converges to an invariant set that over-approximates the Nash equilibrium set.

The topology of value function sets has also garnered interest in the reinforcement learning community (Bellemare, et al., 2019; Dadashi et al., 2019). In Dadashi et al. (2019), the set of value functions generated by policy uncertainty is shown to be a polytope, and Bellman operator-based methods such as value iteration and policy iteration are shown to converge to the value function polytope.

3. MDP and Bellman operator

We introduce our notation for existing results in MDP literature. Contents from this section are discussed in further detail in(Filar & Vrieze, 2012, Sec. 2.2).

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

We consider a *discounted infinite-horizon MDP* defined by $([S], [A], P, C, \gamma)$, where

(1) $[S]$ denotes the finite set of states.

(2) $[A]$ denotes the finite set of actions. Without loss of generality, assume that every action is admissible from each state $s \in [S]$.
(3) $P \in \mathbb{R}^{S\times SA}$ defines the transition kernel. Each component $P_{s',sa}$ is the probability of arriving in state s' by taking state–action (s, a). The matrix P is column stochastic and element-wise non-negative — i.e.,

$$\begin{aligned} &\sum_{s'\in[S]} P_{s',sa} = 1, \ \forall\, (s,a) \in [S]\times[A], \\ &P_{s',sa} \geq 0, \ \forall\, (s',s,a) \in [S]\times[S]\times[A]. \end{aligned} \tag{1}$$

(4) $C \in \mathbb{R}^{S\times A}$ defines the cost matrix. Each component C_{sa} is the cost of state–action pair $(s, a) \in [S] \times [A]$.
(5) $\gamma \in (0, 1)$ denotes the discount factor.

At each time step t, the decision maker chooses an action a at its current state s. The state–action pair (s, a) induces a probability distribution vector over states $[S]$ as $[P_{1,sa}, P_{2,sa}, \ldots, P_{S,sa}]$. The state–action (s, a) also incurs a cost C_{sa} for the decision maker.

The decision maker chooses actions via a *policy*. We denote a policy as a function $\pi : \mathbb{R}^S \times \mathbb{R}^A \to [0, 1]$, where $\pi(s, a)$ denotes the probability that action a is chosen at state s. The set of all policies of an MDP is denoted by Π. Within Π, a policy π is deterministic if at each state s, $\pi(s, a)$ returns 1 for exactly one action, and 0 for all other possible actions. A policy $\pi \in \Pi$ that is not deterministic is a *mixed policy*.

We denote the policy matrix induced by the policy π as $M_\pi \in \mathbb{R}^{S\times SA}$, where

$$(M_\pi)_{s',sa} = \begin{cases} \pi(s,a) & s' = s \\ 0 & s' \neq s. \end{cases} \tag{2}$$

Every policy induces a *Markov chain* (El Chamie, Yu, Açıkmeşe, & Ono, 2018), given by $M_\pi P^\top \in \mathbb{R}^{S\times S}$. Each policy also induces a stationary cost given by

$$\nu(\pi) = \sum_{i\in[S]} e_i e_i^\top M_\pi (\mathbf{1}_S \otimes I_A) C^\top e_i, \ \ \nu(\pi) \in \mathbb{R}^S, \tag{3}$$

where $e_i \in \mathbb{R}^S$ is the unit vector pointing in the ith coordinate, $\otimes$ is the Kronecker product, and I_A is the identity matrix of size A.

For an MDP $([S], [A], P, C, \gamma)$, we are interested in minimizing the *discounted infinite horizon expected cost*, defined as

$$V_s^\star = \min_{\pi\in\Pi} \mathbb{E}_s^\pi \left\{ \sum_{t=0}^{\infty} \gamma^t C_{s^t a^t} \right\}, \quad \forall\, s \in [S], \tag{4}$$

where $\mathbb{E}_s^\pi(f)$ is the discounted infinite horizon expected value of objective f with respect to policy π, s^t and a^t are the state and action taken at time step t, and s is the initial state of the decision maker at $t = 0$.

$V_s^\star$ is the *optimal value function* for the initial state s. The policy $\pi^\star$ that achieves this optimal value is called an *optimal policy*. In general, the optimal value function $V_s^\star$ is unique while the optimal policy $\pi^\star$ is not. The set of optimal policies always includes at least one deterministic stationary policy in the unconstrained setting (Puterman, 2014, Thm. 6.2.11). If there are constraints on the policy and state space, deterministic optimal policies may become infeasible (El Chamie et al., 2018).

3.1. Bellman operator

Determining the optimal value function of a given MDP is equivalent to finding the fixed point of the associated Bellman operator, for which a myriad of techniques exists (Puterman, 2014). We introduce the Bellman operator and its fixed point here for the corresponding MDP problem.

Definition 1 (*Bellman Operator*)**.** For a discounted infinite horizon MDP $([S], [A], P, C, \gamma)$, its Bellman operator $f_C : \mathbb{R}^S \to \mathbb{R}^S$ is given component-wise as

$$\Big(f_C(V)\Big)_s := \min_a \ C_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa} V_{s'}, \ \ \forall\, s \in [S]. \tag{5}$$

The fixed point of the Bellman operator is a value function $V \in \mathbb{R}^S$ that is invariant with respect to the operator.

Definition 2 (*Fixed Point*)**.** $V^\star$ is a fixed point of an operator $F : \mathcal{X} \mapsto \mathcal{X}$ iff

$$V^\star = F(V^\star). \tag{6}$$

In our discussion of the fixed point of the Bellman operator, we consider the following operator properties.

Definition 3 (*Order Preservation*)**.** Let $\mathcal{X}$ be a partially ordered space with partial order $\preceq$. An operator $F : \mathcal{X} \to \mathcal{X}$ is an order preserving operator iff

$$x \preceq x' \to \ F(x) \preceq F(x'), \quad \forall\, x, x' \in \mathcal{X}.$$

Definition 4 (*Contraction*)**.** Let $(\mathcal{X}, d)$ be a complete metric space with metric d. An operator $F : \mathcal{X} \mapsto \mathcal{X}$ is a contracting operator iff

$$d(F(x), F(x')) < d(x, x'), \quad \forall\, x, \ x' \ \in \mathcal{X}.$$

The Bellman operator f_C is known to have both properties on the complete metric space $(\mathbb{R}^S, \|\cdot\|_\infty)$. Therefore, the Banach fixed point theorem can be used to show that f_C has a unique fixed point (Puterman, 2014, Thm. 6.2.3). Because the optimal value function $V^\star$ is given by the unique fixed point of the associated Bellman operator f_C, we use the terms optimal value function and fixed point of f_C interchangeably.

In addition to obtaining $V^\star$, MDPs are also solved to determine the *optimal policy* $\pi^\star$. Every policy π induces a unique *stationary* value function V which satisfies

$$V = \nu(\pi) + \gamma M_\pi P^\top V, \tag{7}$$

where $\gamma \in (0, 1)$. We note that V is a linear function of C through $\nu(\pi)$ as defined in (3), where the dependency is made implicit to simplify notation.

Given a policy π, we can equivalently solve for the stationary value function V as $V = (I - \gamma M_\pi P^\top)^{-1}\nu(\pi)$. From this perspective, the optimal value function $V^\star$ is the minimum vector in $\|\|_\infty$ among the set of stationary value functions corresponding to the set of policies Π. Policy iteration algorithms utilize this fact to obtain the optimal value function $V^\star$ by iterating over the feasible policy space (Puterman, 2014, Sec. 6.4).

Given an input value function V, we can also derive a deterministic optimal policy π associated with $f_C(V)$ as

$$\pi(s,a) := \begin{cases} 1 & a = \underset{a'\in[A]}{\operatorname{argmin}} \ C_{sa'} + \gamma \sum\limits_{s'\in[S]} P_{s',sa'} V_{s'} \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where $\operatorname{argmin}_{a'\in[A]}$ returns the first optimal action a' if multiple actions minimize the expression $C_{sa'} + \gamma \sum_{s'\in[S]} P_{s',sa'} V_{s'}$ at state s.

While policies that solve $f_C(V)$ may not be unique, deterministic or stationary, the policy π derived from (8) will always be unique, deterministic and stationary for a given ordering of actions within the action set. For the remaining sections, we assume that the action set $[A]$ has a fixed ordering of actions.

3.2. Termination criteria for value iteration

Among the many algorithms that solve for the fixed point of the Bellman operator, *value iteration* (VI) is a commonly used and simple technique in which the Bellman operator is iteratively applied until the optimal value function is reached — i.e. starting from any value function $V^0 \in \mathbb{R}^S$, we apply

$$\begin{aligned} V_s^{k+1} &= f_C(V^k) \\ &= \min_{a\in[A]} C_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa} V_{s'}^k, \quad k = 1, 2, \ldots. \end{aligned} \tag{9}$$

The recursion given by (9) converges to the optimal value function of the corresponding discounted infinite horizon MDP. The following result presents a *stopping criteria* for (9).

Lemma 1 (*Puterman, 2014, Thm. 6.3.1*)**.** *For any initial value function $V^0 \in \mathbb{R}^S$, let $\{V^k\}_{k\in\mathbb{N}}$ satisfy the value iteration given by (9). For $\epsilon > 0$, if*

$$\left\|V^{k+1} - V^k\right\|_\infty < \epsilon \frac{(1-\gamma)}{2\gamma},$$

then V^{k+1} is within $\epsilon/2$ of the fixed point $V^\star$, i.e.

$$\left\|V^{k+1} - V^\star\right\|_\infty < \frac{\epsilon}{2}.$$

Lemma 1 connects the relative convergence of sequence $\{V^k\}_{k\in\mathbb{N}}$ to the absolute convergence towards $V^\star$ by showing that the former implies the latter. In general, the stopping criteria differ for different MDP objectives (see Haddad & Monmege, 2018 for recent results on stopping criteria for MDP with a reachability objective).

4. Set-based Bellman operator

The classic Bellman operator with respect to a cost C is well-studied. Motivated by parameter-uncertain MDPs and stochastic games, we extend the classic Bellman operator by *lifting* it to operate on sets rather than individual value function vectors in $\mathbb{R}^S$. For the set-based operator, we analyse its set-based domain and prove relevant operator properties such as order preservation and contraction. Finally, we show the existence of a unique fixed point *set* $\mathcal{V}^\star$ and relate its properties to the fixed point of the classic Bellman operator.

4.1. Set-based operator properties

For the domain of our set-based operator, we define a new metric space $(H(\mathbb{R}^S), d_H)$ based on the Banach space $(\mathbb{R}^S, \|\cdot\|_\infty)$ (Rudin et al., 1964), where $H(\mathbb{R}^S)$ denotes the collection of non-empty compact subsets of $\mathbb{R}^S$. We equip $H(\mathbb{R}^S)$ with *partial order*, $\preceq$, where for $\mathcal{V}, \mathcal{V}' \in H(\mathbb{R}^S)$, $\mathcal{V} \preceq \mathcal{V}'$ iff $\mathcal{V} \subseteq \mathcal{V}'$. The metric d_H is the following *Hausdorff distance* (Henrikson, 1999) defined as

$$\begin{aligned} d_H(\mathcal{V}, \mathcal{V}') := \max\{ &\sup_{V\in\mathcal{V}} \inf_{V'\in\mathcal{V}'} \left\|V - V'\right\|_\infty, \\ &\sup_{V'\in\mathcal{V}'} \inf_{V\in\mathcal{V}} \left\|V - V'\right\|_\infty \}. \end{aligned} \tag{10}$$

Lemma 2 (*Henrikson, 1999, Thm 3.3*)**.** *If $\mathcal{X}$ is a complete metric space, then its induced Hausdorff metric space $(H(\mathcal{X}), d_H)$ is a complete metric space.*

From Lemma 2, since $(\mathbb{R}^S, \|\cdot\|_\infty)$ is a complete metric space, $H(\mathbb{R}^S)$ is a complete metric space with respect to d_H. On the complete metric space $H(\mathbb{R}^S)$, we define a *set-based Bellman operator* which acts on compact sets.

Definition 5 (*Set-based Bellman Operator*)**.** For a family of MDP problems, $([S], [A], P, \mathcal{C}, \gamma)$, where $\mathcal{C} \subset \mathbb{R}^{S\times A}$ is a non-empty compact set, its associated set-based Bellman operator is given by

$$F_{\mathcal{C}}(\mathcal{V}) := \mathrm{cl} \bigcup_{(C,V)\in\mathcal{C}\times\mathcal{V}} f_C(V), \quad \forall\, \mathcal{V} \in H(\mathbb{R}^S),$$

where cl is the closure operator.

Since $F_{\mathcal{C}}$ is the union of uncountably many singleton sets, the resulting set may not be bounded. Therefore, it is not immediately obvious that $F_{\mathcal{C}}(\mathcal{V})$ maps into the metric space $H(\mathbb{R}^S)$.

Proposition 1. *If $\mathcal{C}$ is non-empty and compact, then $F_{\mathcal{C}}(\mathcal{V}) \in H(\mathbb{R}^S)$, $\forall\, \mathcal{V} \in H(\mathbb{R}^S)$.*

Proof. For a non-empty and bounded subset $\mathcal{A}$ of a finite-dimensional real vector space, we define its diameter as $\mathrm{diam}\,(\mathcal{A}) = \sup_{x,y\in\mathcal{A}} \|x - y\|_\infty$. The diameter of a set in a metric space is finite if and only if it is bounded (Rudin et al., 1964).

Take any non-empty compact set $\mathcal{V} \in H(\mathbb{R}^S)$. As $F_{\mathcal{C}}(\mathcal{V}) \subseteq \mathbb{R}^S$, it suffices to prove that $F_{\mathcal{C}}(\mathcal{V})$ is closed and bounded. The closedness is guaranteed by the closure operator. A subset of a metric space is bounded iff its closure is bounded. Hence, to prove the boundedness, it suffices to prove that $\mathrm{diam}\left(\cup_{(C,V)\in\mathcal{C}\times\mathcal{V}} f_C(V)\right) < +\infty$. For any two cost-value function pairs $(C, V), (C', V') \in \mathcal{C} \times \mathcal{V}$,

$$f_C(V) - f_{C'}(V') = \Big(f_C(V) - f_{C'}(V)\Big) + \Big(f_{C'}(V) - f_{C'}(V')\Big). \tag{11}$$

We bound (11) by bounding each of the two terms on the right hand side separately. Due to contraction properties of $f_{C'}$, the second term on the right hand side satisfies $\left\|f_{C'}(V) - f_{C'}(V')\right\|_\infty \leq \gamma \left\|V - V'\right\|_\infty$. To bound the first term, we note that for any two vectors $a, b \in \mathbb{R}^S$,

$$\|a - b\|_\infty = \max\Big\{\max(a - b), \max(b - a)\Big\}, \tag{12}$$

where the operator $\max\{\ldots\}$ returns the maximum element, and $\max(a)$ returns maximum component of vector a. Evaluating $f_{C'}(V) - f_C(V)$ with (12),

$$\begin{aligned} &\max(f_{C'}(V) - f_C(V)) \\ &\leq \max(\nu'(\pi) + \gamma M_\pi P^\top V - \nu(\pi) - \gamma M_\pi P^\top V) \\ &\leq \max\Big(\nu'(\pi) - \nu(\pi)\Big) \\ &\leq \sum_{i\in[S]} \left\|e_i^\top\right\|_\infty \|M_\pi\|_\infty \left\|\mathbf{1}_S \otimes I_A\right\|_\infty \left\|(C' - C)^\top\right\|_\infty \|e_i\|_\infty^2, \end{aligned}$$

where π is an optimal policy corresponding to f_C. Since $\|\mathbf{1}_S \otimes I_A\|_\infty = \|e_i\|_\infty = \left\|e_i^\top\right\|_\infty = \|M_\pi\|_\infty = 1$ for any $\pi \in \Pi$, $\max(f_{C'}(V) - f_C(V)) \leq S \left\|(C' - C)^\top\right\|_\infty$. Similarly, we can show $\max(f_C(V) - f_{C'}(V)) \leq S \left\|(C' - C)^\top\right\|_\infty$. Finally it follows from (11) that

$$\left\|f_C(V) - f_{C'}(V')\right\|_\infty \leq S \left\|(C' - C)^\top\right\|_\infty + \gamma \left\|V - V'\right\|_\infty. \tag{13}$$

Since (13) holds for all $(C, V), (C', V') \in \mathcal{C} \times \mathcal{V}$, and furthermore, for all $C, C' \in \mathcal{C}$ and $V, V' \in \mathcal{V}$,

$$\left\|(C' - C)^\top\right\|_\infty \leq \mathrm{diam}\left(\mathcal{C}^\top\right), \quad \left\|V - V'\right\|_\infty \leq \mathrm{diam}\,(\mathcal{V}),$$

the inequality $\mathrm{diam}\left(\cup_{(C,V)\in\mathcal{C}\times\mathcal{V}} f_C(V)\right) \leq S\,\mathrm{diam}\left(\mathcal{C}^\top\right) + \gamma\,\mathrm{diam}\,(\mathcal{V}) < +\infty$ holds as both $\mathcal{C}^\top$ and $\mathcal{V}$ are bounded. □

Proposition 1 shows that $F_{\mathcal{C}}$ is an operator from $H(\mathbb{R}^S)$ to $H(\mathbb{R}^S)$. Having established the space on which $F_{\mathcal{C}}$ operates, we can draw many parallels between $F_{\mathcal{C}}$ and f_C. Similar to f_C having a unique fixed point $V^\star$ in the vector space, does $F_{\mathcal{C}}$ have a unique *fixed point set* $\mathcal{V}^\star$ which satisfies $F_{\mathcal{C}}(\mathcal{V}^\star) = \mathcal{V}^\star$? To take

the comparison further, since $V^\star$ is optimal for an MDP problem defined by $([S], [A], P, C, \gamma)$, we consider if $\mathcal{V}^\star$ correlates to the *family* of optimal value functions that correspond to the MDP family $([S], [A], P, \mathcal{C}, \gamma)$. We explore these parallels in this paper and derive sufficient conditions for the existence and uniqueness of the fixed point of the set-based Bellman operator $F_{\mathcal{C}}$.

We demonstrate the existence and uniqueness of $\mathcal{V}^\star$ by utilizing the Banach fixed point theorem (Puterman, 2014), which states that a unique fixed point must exist for all contraction operators on a complete metric space. First, we show that $F_{\mathcal{C}}$ has properties given in Definitions 3 and 4 on the complete metric space $(H(\mathbb{R}^S), d_H)$.

Proposition 2. *For any $\mathcal{V} \in H(\mathbb{R}^S)$ and $\mathcal{C} \subset \mathbb{R}^{S\times A}$ closed and bounded, $F_{\mathcal{C}}$ is an order preserving and a contracting operator in the Hausdorff distance.*

Proof. Consider $\mathcal{V}, \mathcal{V}' \in H(\mathbb{R}^S)$ which satisfy $\mathcal{V} \subseteq \mathcal{V}'$, then

$$F_{\mathcal{C}}(\mathcal{V}) = \mathrm{cl} \bigcup_{\substack{(C,V)\\ \in \mathcal{C}\times\mathcal{V}}} f_C(V) \subseteq \mathrm{cl} \bigcup_{\substack{(C,V')\\ \in \mathcal{C}\times\mathcal{V}'}} f_C(V') = F_{\mathcal{C}}(\mathcal{V}').$$

We conclude that $F_{\mathcal{C}}$ is order-preserving. To see that $F_{\mathcal{C}}$ is contracting, we need to show

$$\sup_{V\in F_{\mathcal{C}}(\mathcal{V})} \inf_{V'\in F_{\mathcal{C}}(\mathcal{V}')} \left\|V - V'\right\|_\infty < d_H(\mathcal{V}, \mathcal{V}') \tag{14}$$

$$\sup_{V'\in F_{\mathcal{C}}(\mathcal{V}')} \inf_{V\in F_{\mathcal{C}}(\mathcal{V})} \left\|V - V'\right\|_\infty < d_H(\mathcal{V}, \mathcal{V}'), \tag{15}$$

First we note that taking sup (inf) of a continuous function over the closure of a set $\mathcal{A}$ is equivalent to taking the sup (inf) over $\mathcal{A}$ itself. Furthermore, the single-cost Bellman operator $f_C(V)$ is an element of the set-based Bellman operator $\cup_{(C,V)\in\mathcal{C}\times\mathcal{V}} f_C(V)$ iff $(C, V) \in \mathcal{C} \times \mathcal{V}$. Therefore taking the sup(inf) of $\left\|V - V'\right\|_\infty$ over $V \in F_{\mathcal{C}}(\mathcal{V})$ is equivalent to taking the sup(inf) of $\left\|f_C(V) - f_{C'}(V')\right\|_\infty$ over $(C, V) \in \mathcal{C} \times \mathcal{V}$.

Given $\mathcal{V}, \mathcal{V}' \in H(\mathbb{R}^S)$ and for arbitrary $V \in \mathcal{V}$, $C \in \mathcal{C}$,

$$\inf_{\substack{(C',V')\\ \in\mathcal{C}\times\mathcal{V}'}} \left\|f_C(V) - f_{C'}(V')\right\|_\infty \tag{16a}$$

$$\leq \inf_{\substack{(C',V')\\ \in\mathcal{C}\times\mathcal{V}'}} S \left\|(C - C')^\top\right\|_\infty + \gamma \left\|V - V'\right\|_\infty \tag{16b}$$

$$\leq S \left\|(C' - C')^\top\right\|_\infty + \inf_{V'\in\mathcal{V}'} \gamma \left\|V - V'\right\|_\infty \tag{16c}$$

$$\leq \gamma \inf_{V'\in\mathcal{V}'} \left\|V - V'\right\|_\infty, \tag{16d}$$

where in (16b) we take the upper bound derived in (13). In (16c) we haven chosen the matrix $C = C'$ to minimize $\left\|(C - C')^\top\right\|_\infty$. This eliminates the cost term and we arrive at (16d).

Then (14) and (15) simplify to

$$\sup_{V\in F_{\mathcal{C}}(\mathcal{V})} \inf_{V'\in F_{\mathcal{C}}(\mathcal{V}')} \left\|V - V'\right\|_\infty \leq \gamma \sup_{V\in\mathcal{V}} \inf_{V'\in\mathcal{V}'} \left\|V - V'\right\|_\infty,$$

and

$$\sup_{V'\in F_{\mathcal{C}}(\mathcal{V}')} \inf_{V\in F_{\mathcal{C}}(\mathcal{V})} \left\|V - V'\right\|_\infty \leq \gamma \sup_{V'\in\mathcal{V}'} \inf_{V\in\mathcal{V}} \left\|V - V'\right\|_\infty.$$

Therefore $d_H(F_{\mathcal{C}}(\mathcal{V}), F_{\mathcal{C}}(\mathcal{V}')) \leq \gamma d_H(\mathcal{V}, \mathcal{V}')$. Since $\gamma \in (0, 1)$, $F_{\mathcal{C}}$ is a contracting operator on $H(\mathbb{R}^S)$. □

The contraction property of $F_{\mathcal{C}}$ implies that any repeated application of the operator to a set $\mathcal{V}^0 \in H(\mathbb{R}^S)$ results in a sequence of sets such that the consecutive sets become increasingly closer in the Hausdorff distance. It is then natural to consider whether there is a unique set which all $F_{\mathcal{C}}(\mathcal{V}^k)$ converge to.

Theorem 1. *There exists a unique fixed point $\mathcal{V}^\star$ of the set-based Bellman operator $F_{\mathcal{C}}$ as defined in Definition 1, such that $F_{\mathcal{C}}(\mathcal{V}^\star) = \mathcal{V}^\star$, and $\mathcal{V}^\star$ is a closed and bounded set in $\mathbb{R}^S$.*

Furthermore, for any set $\mathcal{V}^0 \in H(\mathbb{R}^S)$, the iteration

$$\mathcal{V}^{k+1} = F_{\mathcal{C}}(\mathcal{V}^k), \tag{17}$$

converges in the Haussdorf distance — i.e.,

$$\lim_{k\to\infty} d_H(F_{\mathcal{C}}(\mathcal{V}^k), \mathcal{V}^\star) = 0.$$

Proof. As shown in Proposition 2, $F_{\mathcal{C}}$ is a contracting operator. From the Banach fixed point theorem (Puterman, 2014, Thm. 6.2.3), there exists a unique fixed point $\mathcal{V}^\star$, and any arbitrary $\mathcal{V}^0 \in H(\mathbb{R}^S)$ will generate a sequence of sets $\{F_{\mathcal{C}}(\mathcal{V}^k)\}_{k\in\mathbb{N}}$ that converges to $\mathcal{V}^\star$. □

4.2. Properties of the fixed point set

For the Bellman operator f_C on the metric space $\mathbb{R}^S$, the fixed point $V^\star$ corresponds to the optimal value function of the MDP associated with cost C. Because there is no direct association of an MDP problem with a set of cost parameters $\mathcal{C}$, we cannot claim the same for the set-based Bellman operator and $\mathcal{V}^\star$. However, $\mathcal{V}^\star$ does have many desirable properties on $H(\mathbb{R}^S)$, especially in terms of set-based value iteration (17) and in connection to the Bellman operator f_C.

We consider the following generalization of value iteration: instead of a fixed cost parameter, we have at each iteration k, a C^k that is randomly chosen from the compact set of cost parameters $\mathcal{C}$. In general, $\lim_{k\to\infty} f_{C^k}(V^k)$ may not exist. However, we can infer from Theorem 1 that the sequence $\{V^k\}$ converges to the set $\mathcal{V}^\star$ in the Hausdorff distance.

Proposition 3. *Let $\{C^k\}_{k\in\mathbb{N}} \subseteq \mathcal{C}$ be a sequence of costs in $\mathcal{C}$, where $\mathcal{C}$ is a compact set within $\mathbb{R}^{S\times A}$. Let us define the iteration*

$$V^{k+1} = f_{C^k}(V^k),$$

for any $V^0 \in \mathbb{R}^S$. Then the sequence $\{V^k\}_{k\in\mathbb{N}}$ satisfies

$$\lim_{k\to\infty} \inf_{V\in\mathcal{V}^\star} \left\|f_{C^k}(V^k) - V\right\|_\infty = 0,$$

where $\mathcal{V}^\star$ is the unique fixed point set of the operator $F_{\mathcal{C}}$.

Proof. Define $\mathcal{V}^0 = \{V^0\}$, then from Definitions 5 and 1, $V^{k+1} = f_{C^k}(V^k) \in F_{\mathcal{C}}(\mathcal{V}^k)$ for all $k \geq 0$.

At each iteration k, we write $\mathcal{V}^{k+1} = F_{\mathcal{C}}(\mathcal{V}^k)$. From Theorem 1, $\mathcal{V}^k$ converges to $\mathcal{V}^\star$ in the Hausdorff distance, $\lim_{k\to\infty} d_H(\mathcal{V}^k, \mathcal{V}^\star) = 0$. Therefore, for every $\delta > 0$, there exists K such that for all $k \geq K$, $d_H(\mathcal{V}^k, \mathcal{V}^\star) < \delta$. Since $f_{C^k}(V^k) \in \mathcal{V}^{k+1}$, $\inf_{V\in\mathcal{V}^\star} \left\|f_{C^k}(V^k) - V\right\|_\infty \leq d_H(\mathcal{V}^{k+1}, \mathcal{V}^\star) < \delta$ must also be true for all $k \geq K$. Therefore $\lim_{k\to\infty} \inf_{V\in\mathcal{V}^\star} \left\|f_{C^k}(V^k) - V\right\|_\infty = 0$. □

Proposition 3 implies that regardless of whether or not the sequence $\{f_{C^k}(V^k)\}_{k\in\mathbb{N}}$ converges, the sequence $\{V^k\}$ must become arbitrarily close in Hausdorff distance to the set $\mathcal{V}^\star$. This has important implications in the stochastic game setting that are further explored in Section 5. On the other hand, Proposition 3 implies that if the sequence $\{V^k\}$ does converge, its limit point must be an element of $\mathcal{V}^\star$.

Corollary 1. *Define the set of fixed points of f_C for each $C \in \mathcal{C}$ as*

$$\mathcal{U} = \bigcup_{C\in\mathcal{C}} \{V \in \mathbb{R}^S \mid f_C(V) = V\},$$

i.e., $\mathcal{U}$ is the set of optimal value functions for the set of MDPs $([S], [A], P, C, \gamma)$ where $C \in \mathcal{C}$. Furthermore, consider all sequences

$\{C^k\}_{k\in\mathbb{N}} \subseteq \mathcal{C}$ such that for $V^0 \in \mathbb{R}^S$, the iteration $V^{k+1} = f_{C^k}(V^k)$ approaches $V = \lim_{k\to\infty} V^k$, and define the set of all such limits as

$$\mathcal{W} = \bigcup_{\{C^k\}_{k\in\mathbb{N}}\subseteq\mathcal{C}} \{V \in \mathbb{R}^S \mid \lim_{k\to\infty} f_{C^k}(V^k) = V, \text{ where } V^0 \in \mathbb{R}^S, V^{k+1} = f_{C^k}(V^k),\ k = 0, 1, \dots\}, \tag{18}$$

then $\mathcal{U} \subseteq \mathcal{W} \subseteq \mathcal{V}^\star$.

Proof. For any $V \in \mathcal{W}$ and $V^\star \in \mathcal{V}^\star$,

$$\left\|V^\star - V\right\|_\infty \leq \left\|V^\star - f_{C^k}(V^k)\right\|_\infty + \left\|f_{C^k}(V^k) - V\right\|_\infty$$

is satisfied for all $k \in \mathbb{N}$. Furthermore, by assumption, each $V \in \mathcal{W}$ has an associated iteration $V^{k+1} = f_{C^k}(V^k)$ whose limit point is equal to V, i.e. $\lim_{k\to\infty} \left\|f_{C^k}(V^k) - V\right\|_\infty = 0$. Additionally,

$$\lim_{k\to\infty} \inf_{V^\star\in\mathcal{V}^\star} \left\|f_{C^k}(V^k) - V^\star\right\|_\infty = 0$$

follows from Proposition 3. Therefore,

$$\inf_{V^\star\in\mathcal{V}^\star} \left\|V^\star - V\right\|_\infty \leq 0, \quad \forall\, V \in \mathcal{W}.$$

Since the infimum over a compact set is always achieved by an element of the set (Rudin et al., 1964), $V = V^\star \in \mathcal{V}^\star$. Therefore $\mathcal{W} \subseteq \mathcal{V}^\star$. To see that $\mathcal{U} \subseteq \mathcal{W}$, take $C^k = C$ for all $k = 0, 1, \dots$, therefore $\mathcal{U} \subseteq \mathcal{W}$. □

Remark 1. We make the distinction between $\mathcal{V}^\star$, $\mathcal{W}$, and $\mathcal{U}$ to emphasize that $\mathcal{V}^\star$ is not simply the set of fixed points corresponding to f_C for all possible $C \in \mathcal{C}$, given by $\mathcal{U}$, or all the feasible limits of f_{C^k} for some sequence $\{C^k\}_{k\in\mathbb{N}} \subset \mathcal{C}$, given by $\mathcal{W}$. The fixed point set $\mathcal{V}^\star$ contains all possible limiting trajectories of $\{f_{C^k}(V^k)\}_{k\in\mathbb{N}}$ without assuming a limit exists.

In Corollary 1, $\mathcal{U}$ can be easily understood as the set of optimal value functions for the set of standard MDPs $([S], [A], P, C, \gamma)$ generated by $C \in \mathcal{C}$. An interpretation for $\mathcal{W}$ is perhaps less obvious. We use the following example to illustrate the differences between these three sets.

Example 1. Consider a single state, two action MDP with a discount factor $\gamma = 0.9$, where $\mathcal{C}$ is given by $\{\begin{bmatrix}0 & 1\end{bmatrix}, \begin{bmatrix}0 & 2\end{bmatrix}, \begin{bmatrix}1 & 1\end{bmatrix}\}$. Here, $\mathcal{U} = \{0, 10\}$ corresponds to the three optimal value functions when cost is fixed —i.e., where $C^k = C \in \mathcal{C}$. We note that if $\{C^k\} \subseteq \{\begin{bmatrix}0 & 1\end{bmatrix}, \begin{bmatrix}0 & 2\end{bmatrix}\}$, then $V^\star = 0$ regardless of how C^k is chosen. Therefore $\mathcal{W} = \{0\} \cup \mathcal{U} = \mathcal{U}$. Finally, if C^k is randomly chosen from $\mathcal{C}$ and $V^0 = 0$, V^k will randomly fluctuate but satisfy $V^k \in \mathcal{V}^\star = [0, 10]$.

In the context of robust MDPs, $\mathcal{U}$ contains all the fixed point value functions of regular MDPs. The value function set $\mathcal{W}$ contains the fixed point value functions which are *invariant* to fluctuating costs within a subset of $\mathcal{C}$. On the other hand, even if the value functions do not converge, the value function trajectory will still converge to $\mathcal{V}^\star$. Therefore if the goal is to bound the asymptotic behaviour of V^k, it is more useful to determine $\mathcal{V}^\star$.

We summarize our results on the set-based Bellman operator as follows: given a compact set of cost parameters $\mathcal{C}$, $F_\mathcal{C}$ converges to a unique compact set $\mathcal{V}^\star$. The set $\mathcal{V}^\star$ contains all the fixed points of f_C for $C \in \mathcal{C}$. Furthermore, $\mathcal{V}^\star$ also contains the limits of $f_{C^k}(V^k)$ for any $\{C^k\}_{k\in\mathbb{N}} \subseteq \mathcal{C}$, $V^0 \in \mathbb{R}^S$, given that $\lim_{k\to\infty} V^k$ converges. Even if the limit does not exist, V^k must asymptotically converge to $\mathcal{V}^\star$ in the Hausdorff distance.

5. Single-controller stochastic games

In this section, we further elaborate on the properties of the fixed point set $\mathcal{V}^\star$ in the context of single-controller stochastic games, and show that with an appropriate over-approximation of the Nash equilibria cost parameters, $\mathcal{V}^\star$ contains the optimal value functions for player one at Nash equilibria.

A *stochastic game* extends a standard MDP to the competitive multi-agent setting (Shapley, 1953). In the interest of clarity, we define Nash equilibria as well as player value functions in the context of two-player stochastic games. However, the following definitions extend to the N-player stochastic game scenario (Filar & Vrieze, 2012).

We note that the stochastic game we discuss here implicitly assumes *imperfect information* (Filar & Vrieze, 2012, Def. 6.3.6) — at every state, both players have multiple actions to choose from. Therefore, each player's choice of action induces uncertainty in their opponent's costs.

In a two-player stochastic game, both players solve their own MDP while sharing the same states and dynamics. As opposed to standard MDPs, each player's cost and transition kernel depend on the *joint policy*, $\pi = (\pi_1, \pi_2)$, where π_1 and π_2 are respectively player one and player two's policies as defined for standard MDPs in Section 3. The set of joint policies is given by Π, while player one's and player two's sets of policies are given by Π_1 and Π_2, respectively. We denote the actions of player one by a and the actions of player two by b. Players share a common state space given by $[S]$. The transition kernel of the shared dynamics is determined by the tensor $Q \in \mathbb{R}^{S\times S\times A_1\times A_2}$, where Q satisfies

$$\sum_{s'\in[S]} Q_{s'sab} = 1, \ \forall\, (s, a, b) \in [S] \times [A_1] \times [A_2],$$

$$Q_{s'sab} \geq 0, \ \forall\, (s', s, a, b) \in [S] \times [S] \times [A_1] \times [A_2].$$

Each player's cost is given by $D^i \in \mathbb{R}^{S\times A_1\times A_2}$, where D^1_{sab} and D^2_{sab} denote player one and player two's cost when the joint action (a, b) is taken from state s, respectively.

For a specific policy adopted by player two, player one's transition kernel and cost can be represented using the same notation of Section 3. When player two applies policy π_2, player one's transition kernel is given by

$$P^1(\pi_2) \in \mathbb{R}^{S\times SA_1}, \ P^1_{s',sa}(\pi_2) = \sum_{b\in[A_2]} (\pi_2)_{sb} Q_{s'sab}. \tag{19}$$

Furthermore, player one's cost is given by

$$C^1(\pi_2) \in \mathbb{R}^{S\times A_1}, \ C^1_{sa}(\pi_2) = \sum_{b\in[A_2]} (\pi_2)_{sb} D^1_{sab}. \tag{20}$$

For a specific π_1 adopted by player one, player two's cost $C^2(\pi_1)$ and transition kernel $P^2(\pi_1)$ can be similarly defined. Each player then solves a discounted MDP given by $([S], [A_i], P^i(\pi_j), C^i(\pi_j), \gamma_i)$. Since each player only controls a part of the joint action space, the generalization to the joint action space introduces *non-stationarity* in the transition and cost, when viewed from the perspective of an individual player solving an MDP.

Given a joint policy (π_1, π_2), each player attempts to minimize its value function. Player i's optimal discounted infinite horizon expected cost is given by

$$V^i_s = \min_{\pi_i\in\Pi_i} \mathbb{E}^{\pi_i}_s \left\{\sum_{t=0}^{\infty} \gamma_i^t C^i_{s^t a^t}(\pi_j)\right\}, \quad \forall\, s \in [S]. \tag{21}$$

Given a joint policy $\pi = (\pi_1, \pi_2)$, both players have unique stationary value functions $\left(V^1(\pi_1, \pi_2), V^2(\pi_1, \pi_2)\right)$ given by

$$V^1(\pi_1, \pi_2) = \nu^1(\pi_1, \pi_2) + \gamma_1 M_{\pi_1} P^1(\pi_2)^\top V^1(\pi_1, \pi_2), \tag{22a}$$

$$V^2(\pi_1, \pi_2) = \nu^2(\pi_1, \pi_2) + \gamma_2 M_{\pi_2} P^2(\pi_1)^\top V^2(\pi_1, \pi_2), \tag{22b}$$

where $\nu^1(\pi_1, \pi_2) = \sum_{i\in[S]} e_i e_i^\top M_{\pi_1}(\mathbf{1}_s \otimes I_{A_1}) C^1(\pi_2)^\top e_i$ and $\nu^2(\pi_1, \pi_2) = \sum_{i\in[S]} e_i e_i^\top M_{\pi_2}(\mathbf{1}_s \otimes I_{A_2}) C^2(\pi_1)^\top e_i$. Since a two-player

stochastic game can be viewed as two coupled MDPs, the MDP notion of optimality must be expanded to reflect the dependency of a player's individual optimal policy on the joint policy space. We define a *Nash equilibrium* in terms of each player's value function (Filar & Vrieze, 2012, Sec. 3.1).

Definition 6 (*Two-Player Nash Equilibrium*)**.** A joint policy $\pi^\star = (\pi_1^\star, \pi_2^\star)$ is a Nash equilibrium if the corresponding value functions as given by (22) satisfy

$$V^1(\pi_1^\star, \pi_2^\star) \leq V^1(\pi_1, \pi_2^\star), \quad \forall\, \pi_1 \in \Pi_1,$$

$$V^2(\pi_1^\star, \pi_2^\star) \leq V^2(\pi_1^\star, \pi_2), \quad \forall\, \pi_2 \in \Pi_2.$$

We also denote the Nash equilibrium value functions as $V^1(\pi^\star)$ and $V^2(\pi^\star)$ and the set of Nash equilibria for a stochastic game as $\Pi_{NE} \subset \Pi$.

Definition 6 implies that a Nash equilibrium is achieved when the joint policy simultaneously generates both value functions $V^1(\pi^\star)$ and $V^2(\pi^\star)$, which are the fixed points of the Bellman operator with respect to parameters $\left(C^1(\pi_2), P^1(\pi_2)\right)$ and $\left(C^2(\pi_1), P^2(\pi_1)\right)$, respectively — i.e. $V^1(\pi_1^\star, \pi_2^\star) = \min_{\pi_1 \in \Pi_1} \left\{ v^1(\pi_1, \pi_2^\star) + \gamma_1 M_{\pi_1} P^1(\pi_2^\star)^\top V^1(\pi_1^\star, \pi_2^\star) \right\}$, and $V^2(\pi_1^\star, \pi_2^\star) = \min_{\pi_2 \in \Pi_2} \left\{ v^2(\pi_1^\star, \pi_2) + \gamma_2 M_{\pi_2} P^2(\pi_1^\star)^\top V^2(\pi_1^\star, \pi_2^\star) \right\}$.

In general, a stochastic game does not have a unique Nash equilibrium. Furthermore, Nash equilibria policies are not necessarily composed of deterministic individual policies. Therefore while each player's Nash equilibrium value function is always the fixed point of the associated Bellman operator, the Nash equilibrium policy for each player may *not* be the optimal deterministic policy associated to the Nash equilibrium value function. The existence of at least one Nash equilibrium for any general-sum stochastic game is given in Filar and Vrieze (2012). When the stochastic game is also zero-sum, all Nash equilibria correspond to a unique value function.

Since the technical content of this paper does not address non-stationarity in the transition kernel, we focus on analysing non-stationarity in the cost term. Specifically, we constrain our analysis to a *single-controller stochastic game* (Filar & Vrieze, 2012), i.e. when the transition kernel is controlled by player one only. Single-controller stochastic games form an important class of games to model dynamic control in queueing networks (Altman, 1994) and attacker–defender games with stochastic transitions (Ang, Chan, Jiang, & Yeoh, 2017; Eldosouky, Saad, & Niyato, 2016). Similar to our discussion of a two-player Nash equilibrium, we exclusively consider a two-player single-controller stochastic game. However, we note that the following definition can be extended to an N-player single-controller stochastic game in which the transition kernel is independent of all but one player's actions.

Definition 7 (*Single-controller stochastic game*)**.** A single-controller stochastic game is a two-player stochastic game where the probability transition kernel is independent of player two's actions, i.e., for each $(s', s, a) \in [S] \times [S] \times [A_1]$

$$Q_{s'sab} = Q_{s'sab'}, \quad \forall\, b, b' \in [A_2],$$

i.e. $P^1(\pi_2) = P$, $\forall\, \pi_2 \in \Pi_2$ and $P^2(\pi_1)_{s',sb} = P^2(\pi_1)_{s',sb'}$, $\forall\, b, b' \in [A_2]$, $\pi_1 \in \Pi_1$.

Although both players are still optimizing their value functions in a single-controller stochastic game, player two's policy only affects its immediate cost at each state, while its transition dynamic becomes a time-varying Markov chain. Furthermore, player two's policy affects player one's MDP through cost matrix $C^1(\pi_2)$.

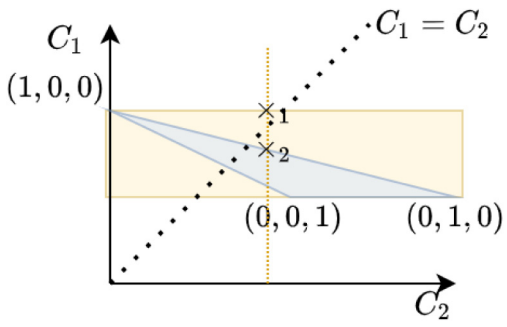


Fig. 1. Feasible player costs vs interval set over-approximation.

We analyse a single-controller game from the set-based MDP perspective by utilizing Proposition 3. Suppose we are given a compact set $\mathcal{C} \subset \mathbb{R}^{S \times A_1}$ that over-approximates $\mathcal{C}^{NE}$, the set of cost parameters observed by player one at Nash equilibria, given by

$$\mathcal{C}^{NE} = \{C^1(\pi_2^\star) \in \mathbb{R}^{S \times A_1} \mid (\pi_1^\star, \pi_2^\star) \in \Pi_{NE}\} \subseteq \mathcal{C}. \tag{23}$$

Then the Nash equilibria value functions belong to the fixed point set of $F_{\mathcal{C}}$. The simplest over-approximations of $\mathcal{C}^{NE}$ is the interval set of all feasible costs.

Example 2 (*Interval Set Approximation*)**.** An approximation to $\mathcal{C}^{NE}$ can always be given by interval sets. At each state–action pair (s, a), the MDP cost parameter for player one is given by (20). We can take the maximum and minimum elements of the set $\{D^1_{sab}\}_{b \in [A_2]}$ for all state actions pairs (s, a) to form an interval set $\mathcal{C} = \mathcal{C}_{11} \times \cdots \times \mathcal{C}_{SA_1} \in H(\mathbb{R}^{S \times A_1})$, such that

$$\mathcal{C}_{sa} = \{D^1_{sab}\}_{b \in [A_2]} = [\underline{C}_{sa}, \overline{C}_{sa}],$$

where $\underline{C}_{sa} = \min_{b \in [A_2]} D^1_{sab}$ and $\overline{C}_{sa} = \max_{b \in [A_2]} D^1_{sab}$ can be directly observed.

Interval set approximation will always give an admissible approximation. However, more general sets such as polytopes allow for more accurate representations of the player's feasible costs.

Example 3 (*Polytope Set Approximation*)**.** Consider the set of costs at a particular state s in a two-player single-controller stochastic game, for which $A_1 = 2$ and $A_2 = 3$. Player one's costs corresponding to player two's deterministic policies are given by points $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ in Fig. 1. Any mixed policy from player two will result in an expected cost for player one that corresponds to a point within the blue region in Fig. 1. On the other hand, the approximation from Example 2 is given by the yellow region. In this example, we can observe that the interval set generously over-approximates player one's feasible costs.

An over-approximation of the set of feasible costs will also over-approximate the possible limiting trajectories for a player's learning algorithm. Consider the costs at point $\times_1 = (C_2', C_1')$ in Fig. 1, here value iteration would have chosen a_2 corresponding to C_2' in its value iteration and return the corresponding value function and transition kernel for state s. However, the *true cost* when action a_2 has equivalent cost C_2' is at $\times_2 = (\bar{C}_1, C_2')$ on the boundary of the blue polytope. Since $\times_2$ lies below the line $C_1 = C_2$, a_1 corresponding to $\bar{C}_1$ is the actual optimal action. Not only does the true cost $(\bar{C}_1, C_2')$ result in a different value function and transition kernel, the value function and transition kernel when costs are (C_2', C_1') are infeasible.

The set of feasible costs itself is an over-approximation of the Nash equilibria cost set $\mathcal{C}^{NE}$. As Example 3 shows, the extension from interval sets to compact sets enables additional information (feasible costs, knowledge of opponents' policy) to be used to approximate $\mathcal{C}^{NE}$ to greater accuracy.

Given a compact set $\mathcal{C}$ that over-approximates the set of player one's cost parameters at Nash equilibria, $\mathcal{C}^{NE}$, we now show that the Nash equilibria value functions for player one must lie within $\mathcal{V}^\star$, the fixed point set of $F_{\mathcal{C}}$.

Theorem 2. *In a single-controller stochastic game, let $\mathcal{C} \subset \mathbb{R}^{S\times A_1}$ be an over-approximation of the Nash equilibria costs formulated in (23). If $\mathcal{C}$ is compact, then the set of stationary value functions for player one at Nash equilibria policies (22a) is a subset of $\mathcal{V}^\star$, the fixed point set of $F_{\mathcal{C}}$.*

Proof. We define the set of Nash equilibria value functions for player one as

$$\mathcal{V}^{NE} = \left\{ V \in \mathbb{R}^S \mid V = f_{C^1(\pi_2^\star)}(V) \right\}, \tag{24}$$

where the Bellman operator $f_{C^1(\pi_2^\star)}$ is defined with P, the π_2-independent transition kernel for both players. For any $V^\star \in \mathcal{V}^{NE}$, there exists $C^\star = C^1(\pi_2^\star) \in \mathcal{C}$ such that $V^\star$ is the fixed point of $f_{C^\star}$. Then from Corollary 1, $V^\star \in \mathcal{V}^\star$. □

Remark 2. Although the Nash equilibrium value function $V^\star$ is always the unique fixed point of $f_{C^\star}$ given by (5), where $C^\star$ is player one's cost at Nash equilibrium, we note that in general, player one's policy at Nash equilibrium is not the optimal deterministic policy of $f_{C^\star}(V^\star)$ given by (8); this is because the joint policy at Nash equilibrium may not be composed of deterministic individual policies, while the solution to (8) is always deterministic.

However, player one's policy at Nash equilibrium must be a convex combination of all deterministic policies that solve (8) (Filar & Vrieze, 2012).

We summarize the application of the set-based MDP framework to single-controller stochastic games as the following: when $\mathcal{C}$ over-approximates the set of costs at Nash equilibria, the fixed point set $\mathcal{V}^\star$ of operator $F_{\mathcal{C}}$ contains all of the Nash equilibria value functions for player one in a single-controller stochastic game.

6. Application to interval set-based Bellman operator

In this section, we show that when the cost parameter set $\mathcal{C}$ and the initial value function set $\mathcal{V}^0$ are interval sets, the fixed point set $\mathcal{V}^\star$ of $F_{\mathcal{C}}$ is also an interval set, as done similarly in Givan et al. (2000). However, we note that convergence in Givan et al. (2000) is shown under an unconventional partial ordering scheme. Leveraging our set-based Bellman operator framework and the Hausdorff distance as our metric, our results are derived in a more straightforward manner using interval arithmetics.

As shown in Examples 2 and 3, interval sets over-approximate the set of Nash equilibria costs given by (23). In this section, we compute the fixed point set $\mathcal{V}^\star$ of an interval set-based Bellman operator. Suppose the set of costs is given by

$$\mathcal{C} = \left\{ C \in \mathbb{R}^{S\times A} \mid C_{sa} \in [\underline{C}_{sa}, \overline{C}_{sa}],\ \forall\, (s,a) \in [S]\times[A] \right\}. \tag{25}$$

and the set of input value functions is given by

$$\mathcal{V} = \left\{ V \in \mathbb{R}^S \mid V_s \in [\underline{V}_s, \overline{V}_s],\ \forall\, s \in [S] \right\}. \tag{26}$$

6.1. Hausdorff distance between interval sets

We first show that the Hausdorff distance between two interval sets $\mathcal{V}, \mathcal{V}' \in H(\mathbb{R}^S)$ can be computed by strictly using the upper and lower bounds of the intervals.

Lemma 3. *For two intervals $\mathcal{X}, \mathcal{Y} \in H(\mathbb{R}^S, \|\cdot\|_\infty)$ given by $\mathcal{X} = [\underline{x}, \overline{x}]$ and $\mathcal{Y} = [\underline{y}, \overline{y}]$, where $\underline{x}, \overline{x}, \underline{y}, \overline{y} \in \mathbb{R}^S$, the Hausdorff distance (10) can be calculated as*

$$d_H(\mathcal{X}, \mathcal{Y}) = \max\{ \left\| \underline{x} - \underline{y} \right\|_\infty, \|\overline{x} - \overline{y}\|_\infty \}.$$

Proof. We consider the component-wise Hausdorff distance by noting that when coupled with the infinity norm on $\mathbb{R}^S$, the Hausdorff distance satisfies

$$d_H(\mathcal{X}, \mathcal{Y}) = \max_{i\in[S]}\ d_H(\mathcal{X}_i, \mathcal{Y}_i),$$

where $\mathcal{X} = \mathcal{X}_1 \times \cdots \mathcal{X}_S$ and $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \mathcal{Y}_S$ (Chavent, 2004).

We first compute $d_H(\mathcal{X}_i, \mathcal{Y}_i)$, where $\mathcal{X}_i = [\underline{x}_i, \overline{x_i}]$ and $\mathcal{Y}_i = [\underline{y}_i, \overline{y_i}]$ are interval sets. Recall that the infinity norm can be written using max operators given in (12). The nested max representation of the infinity norm allows us to directly evaluate the infimum and supremum of $\|x_i - y_i\|_\infty$ over $\mathcal{X}_i$ and $\mathcal{Y}_i$ respectively, as

$$\sup_{y_i\in\mathcal{Y}_i} \inf_{x_i\in\mathcal{X}_i} \|x_i - y_i\|_\infty = \max\{\max(\underline{x}_i - \underline{y}_i), \max(\overline{y}_i - \overline{x}_i)\}.$$

Similarly, we can derive

$$\sup_{x_i\in\mathcal{X}_i} \inf_{y_i\in\mathcal{Y}_i} \|x_i - y_i\|_\infty = \max\{\max(\underline{y}_i - \underline{x}_i), \max(\overline{x}_i - \overline{y}_i)\}.$$

Finally we recall the Hausdorff distance from (10):

$$\begin{aligned} d_H(\mathcal{X}_i, \mathcal{Y}_i) &= \max\{ \sup_{y_i\in\mathcal{Y}_i} \inf_{x_i\in\mathcal{X}_i} \|x_i - y_i\|_\infty, \\ &\qquad \sup_{x_i\in\mathcal{X}_i} \inf_{y_i\in\mathcal{Y}_i} \|x_i - y_i\|_\infty \} \\ &= \max\{\max(\underline{x}_i - \underline{y}_i), \max(\overline{y}_i - \overline{x}_i), \\ &\qquad \max(\underline{y}_i - \underline{x}_i), \max(\overline{x}_i - \overline{y}_i)\} \\ &= \max\{ \left\| \underline{x}_i - \underline{y}_i \right\|_\infty, \|\overline{x}_i - \overline{y}_i\|_\infty \}. \end{aligned} \tag{27}$$

Then the total Hausdorff distance between $\mathcal{X}$ and $\mathcal{Y}$ is given by

$$\begin{aligned} d_H(\mathcal{X}, \mathcal{Y}) &= \max_{i\in[S]}\{\max\{ \left\| \underline{x}_i - \underline{y}_i \right\|_\infty, \|\overline{x}_i - \overline{y}_i\|_\infty \}\} \\ &= \max\{ \left\| \underline{x} - \underline{y} \right\|_\infty, \|\overline{x} - \overline{y}\|_\infty \}. \end{aligned} \tag{28}$$

□

Lemma 3 shows that interval sets are *nice* in that their Hausdorff distances can be derived via component-wise operations on the boundaries of the intervals. We use Lemma 3 later in this section to obtain convergence guarantees of set-based value iteration to the fixed point set of the interval set-based Bellman operator.

6.2. Interval arithmetic

To compute the fixed point of an interval set-based Bellman operator, we introduce some relevant interval arithmetic operators (Moore, 1966).

$$\begin{aligned} \alpha[a,b] &= [\alpha a, \alpha b], \qquad \alpha \geq 0, \\ [a,b] + [c,d] &= [a+c, b+d], \\ [a,b] - [c,d] &= [a-d, b-c], \\ \min\{[a,b],[c,d]\} &= [\min\{a,c\}, \min\{b,d\}], \end{aligned} \tag{29}$$

where the last operator $\min\{[a,b],[c,d]\}$ denotes the smallest interval that contains $\{\min\{x,y\}, \mid x \in [a,b], y \in [c,d]\}$. Since

the last equivalence statement in (29) is not as obvious as the standard addition and subtraction operators, we prove it in the following Lemma.

Lemma 4. *The* min *operator for interval sets can be calculated as*

$$\min\{[a,b],[c,d]\} = [\min\{a,c\}, \min\{b,d\}].$$

Proof. Let us consider the sets $\mathcal{A} = \min\{[a,b],[c,d]\}$ and $\mathcal{B} = [\min\{a,c\}, \min\{b,d\}]$. We first show that $\mathcal{A} \subseteq \mathcal{B}$: for $z \in \min\{[a,b],[c,d]\}$, there exist $x \in [a,b]$ and $y \in [c,d]$ such that $z = \min\{x,y\}$. Then necessarily, $\min\{a,c\} \leq z$ and $z \leq \min\{b,d\}$ must be satisfied.

To prove the inclusion $\mathcal{B} \subseteq \mathcal{A}$, take $v \in [\min\{a,c\}, \min\{b,d\}]$. If $v \in [a,b]$, then $v = \min\{v, \max\{v,d\}\}$. If $\max\{v,d\} = d$, then $v \in \min\{[a,b],[c,d]\}$ follows from $v \in [a,b]$ and $d \in [c,d]$. If $\max\{v,d\} = v$, then $d < v \leq b$. This contradicts $v \in [\min\{a,c\}, \min\{b,d\}]$.

If $v \notin [a,b]$, then either $a \neq \min\{a,c\}$ or $b \neq \min\{b,d\}$. This is equivalent to either $c \leq v < a$ or $b < v \leq d$ being true. $b < v \leq d$ cannot be true since $v \in [\min\{a,c\}, \min\{b,d\}]$. $c \leq v < a$ implies that $v \in [c,d]$ and $v = \min\{v,a\}$, then $v \in \min\{[a,b],[c,d]\}$. □

With Lemmas 3 and 4, we can analytically compute the fixed point set of an interval set-based Bellman operator and give convergence guarantees of interval set-based value iteration.

Proposition 4. *For interval sets $\mathcal{C}$ and $\mathcal{V}$ given by (25) and (26), respectively, $F_{\mathcal{C}}(\mathcal{V})$ as defined in Definition 5 is an interval set and can be formulated as*

$$F_{\mathcal{C}}(\mathcal{V}) = \{V \mid V \leq V_u,\ -V \leq -V_l,\ V \in \mathbb{R}^S\},$$

for $V_l = f_{\underline{C}}(\underline{V})$ and $V_u = f_{\overline{C}}(\overline{V})$.

Furthermore, the sequence $\{\mathcal{V}^k\}_{k\in\mathbb{N}}$ generated by the iteration $\mathcal{V}^{k+1} = F_{\mathcal{C}}(\mathcal{V}^k)$ starting from any interval set $\mathcal{V}^0$ will converge to $\mathcal{V}^\star$ in Hausdorff distance: for every $\epsilon > 0$, there exists $\mathcal{V}^k$ which satisfies

$$d_H(\mathcal{V}^k, \mathcal{V}^\star) \leq \epsilon/2, \tag{30}$$

where (30) is satisfied if $d_H(\mathcal{V}^k, \mathcal{V}^{k-1})\frac{2\gamma}{1-\gamma} < \epsilon$.

Proof. We recall Definition 5 for the set-based Bellman operator and the component-wise definition of $f_{\mathcal{C}}$ in Definition 1. Using these definitions and the fact that $\mathcal{C} = [\underline{C}, \overline{C}]$ and $\mathcal{V} = [\underline{V}, \overline{V}]$ are both interval sets, the set-based Bellman operator can be written as

$$\Big(F_{\mathcal{C}}(\mathcal{V})\Big)_s = \text{cl} \bigcup_{\substack{C\in[\underline{C},\overline{C}]\\ V\in[\underline{V},\overline{V}]}} \min_{a\in[A]} C_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa}V_{s'}.$$

Let $G(C_{sa}, V) = C_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa}V_{s'}$. G be a continuous function and order preserving in its inputs C_{sa} and V. Therefore the union over interval sets in $\big(F_{\mathcal{C}}(\mathcal{V})\big)_s$ can be written using interval arithmetic notation as

$$\bigcup_{\substack{C\in[\underline{C},\overline{C}]\\ V\in[\underline{V},\overline{V}]}} \min_{a\in[A]} C_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa}V_{s'} \tag{31}$$

$$= \left\{ \min_{a\in[A]} C_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa}V_{s'} \mid C \in [\underline{C}, \overline{C}], V \in [\underline{V}, \overline{V}] \right\} \tag{32}$$

$$= \min_{a\in[A]} [\underline{C}_{sa}, \overline{C}_{sa}] + \gamma \sum_{s'\in[S]} P_{s',sa}[\underline{V}_{s'}, \overline{V}_{s'}]. \tag{33}$$

Since interval sets are closed by definition, the closure of (31) must also equal (33). Therefore, $F_{\mathcal{C}}(\mathcal{V})$ can be equivalently written component-wise as

$$\Big(F_{\mathcal{C}}(\mathcal{V})\Big)_s = \min_{a\in[A]} [\underline{C}_{sa}, \overline{C}_{sa}] + \gamma \sum_{s'\in[S]} P_{s',sa}[\underline{V}_{s'}, \overline{V}_{s'}]. \tag{34}$$

Then, $\gamma > 0$ and $P_{s',sa} \geq 0$ for all $(s', s, a) \in [S] \times [S] \times [A]$ allow us to directly perform interval arithmetic component-wise for $F_{\mathcal{C}}$ as

$$\Big(F_{\mathcal{C}}(\mathcal{V})\Big)_s = \min_{a\in[A]} \Big[\underline{C}_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa}\underline{V}_{s'},\ \overline{C}_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa}\overline{V}_{s'} \Big] \tag{35a}$$

$$= \Big[\min_{a\in[A]} \underline{C}_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa}\underline{V}_{s'},\ \min_{a\in[A]} \overline{C}_{sa} + \gamma \sum_{s'\in[S]} P_{s',sa}\overline{V}_{s'} \Big] \tag{35b}$$

$$= [\Big(f_{\underline{C}}(\underline{V})\Big)_s,\ \Big(f_{\overline{C}}(\overline{V})\Big)_s], \tag{35c}$$

where (35b) utilizes the interval set-based minimization derived in Lemma 4, and (35c) follows from Definition 1.

The image of $F_{\mathcal{C}}$ is another closed interval, as shown by (35c). From Theorem 1, any interval set $\mathcal{V}^0 = [\underline{V}, \overline{V}]$ generates an iteration $\mathcal{V}^{k+1} = F_{\mathcal{C}}(\mathcal{V}^k)$ which satisfies $\lim_{k\to\infty} F_{\mathcal{C}}(\mathcal{V}^k) = \mathcal{V}^\star$. We can use interval arithmetic to derive $\mathcal{V}^\star = \lim_{k\to\infty} F_{\mathcal{C}}(\mathcal{V}^k) = \Big[\lim_{k\to\infty} f_{\underline{C}}(\underline{V}^k), \lim_{k\to\infty} f_{\overline{C}}(\overline{V}^k)\Big] = [\underline{V}^\star, \overline{V}^\star]$, where $\underline{V}^\star$ and $\overline{V}^\star$ are the fixed points of $f_{\underline{C}}$ and $f_{\overline{C}}$, respectively.

At each iteration, the Hausdorff distance between $\mathcal{V}^k$ and $\mathcal{V}^\star$ is given by $d_H(\mathcal{V}^k, \mathcal{V}^\star) = d_H\Big([f_{\underline{C}}(\underline{V}^k), f_{\overline{C}}(\overline{V}^k)], [\underline{V}^\star, \overline{V}^\star]\Big)$. Using Lemma 3, $d_H(\mathcal{V}^k, \mathcal{V}^\star)$ is given by

$$\max\Big\{ \big\|f_{\underline{C}}(\underline{V}^k) - \underline{V}^\star\big\|_\infty, \big\|f_{\overline{C}}(\overline{V}^k) - \overline{V}^\star\big\|_\infty \Big\}.$$

Similarly, $d_H(\mathcal{V}^{k+1}, \mathcal{V}^k)$ is given by

$$\max\Big\{ \big\|f_{\underline{C}}(\underline{V}^k) - f_{\underline{C}}(\underline{V}^{k+1})\big\|_\infty, \big\|f_{\overline{C}}(\overline{V}^k) - f_{\overline{C}}(\overline{V}^{k+1})\big\|_\infty \Big\}.$$

From Lemma 1, if $\big\|f_{\underline{C}}(\underline{V}^k) - f_{\underline{C}}(\underline{V}^{k+1})\big\|_\infty < \epsilon\frac{1-\gamma}{2\gamma}$ for some $\epsilon > 0$, then $\big\|f_{\underline{C}}(\underline{V}^k) - \underline{V}^\star\big\|_\infty < \frac{\epsilon}{2}$. Similarly, if $\big\|f_{\overline{C}}(\overline{V}^k) - f_{\overline{C}}(\overline{V}^{k+1})\big\|_\infty < \epsilon\frac{1-\gamma}{2\gamma}$ for some $\epsilon > 0$, then $\big\|f_{\overline{C}}(\overline{V}^k) - \overline{V}^\star\big\|_\infty < \frac{\epsilon}{2}$. Therefore if $\max\Big\{ \big\|f_{\underline{C}}(\underline{V}^k) - \underline{V}^\star\big\|_\infty, \big\|f_{\overline{C}}(\overline{V}^k) - f_{\overline{C}}(\overline{V}^{k+1})\big\|_\infty \Big\} < \epsilon$, then $\max\Big\{ \big\|f_{\underline{C}}(\underline{V}^k) - \underline{V}^\star\big\|_\infty, \big\|f_{\overline{C}}(\overline{V}^k) - \overline{V}^\star\big\|_\infty \Big\} < \frac{\epsilon}{2}$. Since $d_H(\mathcal{V}^{k+1}, \mathcal{V}^k) = \max\Big\{ \big\|f_{\underline{C}}(\underline{V}^k) - \underline{V}^\star\big\|_\infty, \big\|f_{\overline{C}}(\overline{V}^k) - f_{\overline{C}}(\overline{V}^{k+1})\big\|_\infty \Big\}$ and $d_H(\mathcal{V}^{k+1} - \mathcal{V}^\star) = \max\Big\{ \big\|f_{\underline{C}}(\underline{V}^k) - \underline{V}^\star\big\|_\infty, \big\|f_{\overline{C}}(\overline{V}^k) - \overline{V}^\star\big\|_\infty \Big\}$, we conclude that if $d_H(\mathcal{V}^{k+1}, \mathcal{V}^k) < \frac{(1-\gamma)\epsilon}{2\gamma}$, then $d_H(\mathcal{V}^{k+1} - \mathcal{V}^\star) < \frac{\epsilon}{2}$. □

Remark 3. In existing work, V_l is equivalent to the *optimistic value function* in Iyengar (2005) when the transition kernel is known and cost-uncertainty is given by bounded intervals. Furthermore, Proposition 4 specializes interval value iteration from Givan et al. (2000) to cost-uncertainty only and proves stronger convergence results due to this specialization.

While $\mathcal{V}^k$ converges to $\mathcal{V}^\star$ in Hausdorff distance, each $\mathcal{V}^k$ may not over-approximate $\mathcal{V}^\star$. In fact, if $\underline{V}^k > \underline{V}^\star$ for some $k \in \mathbb{N}$, then $\mathcal{V}^\star \subsetneq \mathcal{V}^k$ for all k. Nonetheless, we can still utilize $\mathcal{V}^k$ to obtain an *over-approximation* of $\mathcal{V}^\star$ by using estimate intervals $\tilde{\mathcal{V}}^{k+1} = [f_{\underline{C}}(\underline{V}^k) - \mathbf{1}_S\epsilon, f_{\underline{C}}(\underline{V}^k) + \mathbf{1}_S\epsilon]$.

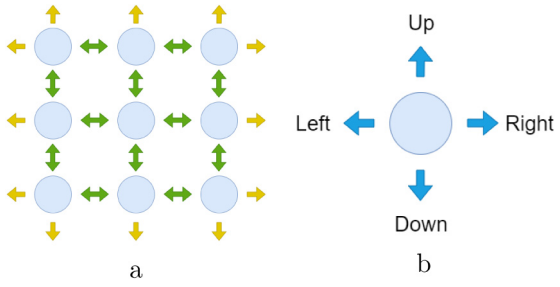


Fig. 2. (a): Each player's state space $[S]$, $S = 9$. Green actions lead to a neighbouring state and yellow actions are infeasible. (b): Actions space $[A]$, $A = 4$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

7. Numerical example

Our set-based Bellman operator framework is motivated by dynamic programming-based learning algorithms in stochastic games. We highlight this connection by applying set-based value iteration in a two-player single-controller stochastic game, and show that both the transient and asymptotic behaviour of player one's value function can be bounded, regardless of the opponent's learning algorithm.

We consider a two-player single-controller stochastic game as defined in Definition 7. Each player solves a discounted MDP given by $([S], [A_{1,2}], P^{1,2}, C^{1,2}, \gamma_{1,2})$ where $A_1 = A_2 = A$ and the players share an identical state–action space $([S], [A])$. Player one's cost is given by

$$C^1_{sa}(\pi_2) = C_{sa} + J_{sb}\pi_2(s, b), \quad \forall\, (s, a) \in [S] \times [A],$$

while player two's cost is given by

$$C^2_{sb}(\pi_1) = C_{sb} - J_{sa}\pi_1(s, a), \quad \forall\, (s, b) \in [S] \times [A].$$

The matrices $C, J \in \mathbb{R}_+^{S\times A}$ are the same for C^1 and C^2.

While there exist algorithms that converge to the Nash equilibrium for such single-controller stochastic games (Hu & Wellman, 2003; Littman, 1994), convergence is not guaranteed if the players do not coordinate on which algorithm to use between themselves. We use the set-based Bellman operator to show that we can determine both the value function set that player one's Nash equilibrium value function belongs to and the value function set that player one's value function trajectory converges to, regardless of what the player two does.

The state space of our stochastic game is a 3×3 grid, shown in Fig. 2(a), where the total number of states is $S = 9$ and the total number of actions per state is $A = 4$. State s' is a neighbouring state of s if it is immediately connected to s by a green arrow in Fig. 2(a). At state s, let $\mathcal{N}_s$ denote the set containing all neighbouring states of s.

As shown in Fig. 2(b), the actions available in each state are labelled 'left', 'right', 'up', or 'down'. From each state s, an action is feasible if it coincides with a green arrow in Fig. 2(a), and infeasible if it coincides with a yellow arrow. For a feasible action a, the target state s' of state–action pair (s, a) is the neighbouring state of s in the direction of the action a. When taking a feasible action a, player one's transition probability is given by

$$P^1_{s'sa} = \begin{cases} 0.7 & s' = \text{ target state} \\ \frac{0.3}{|\mathcal{N}_s|-1} & s' \neq \text{ target state},\ s' \in \mathcal{N}_s \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

If the action a is infeasible, player one's transition probability is given by

$$P^1_{s'sa} = \begin{cases} \frac{1}{|\mathcal{N}_s|} & s' \in \mathcal{N}_s \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

As given in Definition 7, the player two's transition dynamic is independent of player two's actions.

We select matrices $C, J \in \mathbb{R}^{9\times 4}$ by uniformly sampling each component from the interval $[0, 1]$. As in Example 2, the over-approximation of player one's feasible costs as interval sets is given by

$$\mathcal{C} = \Big\{ C^1 \in \mathbb{R}^{9\times 4} \mid C^1_{sa} \in [C_{sa}, C_{sa} + J_{sa}], \\ \forall\, (s, a) \in [9] \times [4] \Big\}, \tag{38}$$

where the upper bound $C_{sa} + J_{sa}$ is achieved when player two's probability of taking action $b = a$ from state s is 1.

We consider the two-player value iteration algorithm that forms the basis for many learning algorithms in stochastic games (Filar & Vrieze, 2012; Littman, 1994), summarized in Algorithm 1. At step k, player one solves for the optimal policy π^{k+1} given by the Bellman operator using (8). Player two obtains its optimal policy using the function $g : \Pi_1 \to \Pi_2$, we do not make any assumptions on g beyond that it is entirely a function of π_1.

Algorithm 1 Two-player value iteration

Input: $([S], [A], P^1, C^1, \gamma_1), V_0$.
Output: $V^\star, \pi_1^\star$
  $\pi_1^0(s) = \pi_2^0(s) = 0, \forall\, s \in [S]$
  for $k = 0, \ldots,$ do
    $C = C^1(\pi_1^k, \pi_2^k)$
    $(V^{k+1}, \pi_1^{k+1}) = f_C(V^k)$
    $\pi_2^{k+1} = g(\pi_1^{k+1})$
  end for

Our analysis provides bounds on player one's value function when we do not know how player two is updating its policy — i.e. when g is unknown. In simulation, we take g to be different strategies and show that player one's value functions are bounded by the interval set analysis and converge towards the fixed point set of the corresponding Bellman operator.

Suppose that both players update their policies via value iteration (8) under different discount factors. Player one performs value iteration with a discount factor of $\gamma_1 = 0.7$, while player two performs value iteration with an unknown discount factor $\gamma_2 \in (0, 1)$. Assuming both players' value functions are initialized to be 0 in every state, we simulate player one's value function trajectories for different values of γ_2 in Fig. 3.

Fig. 3 shows that when player two utilizes different discount factors, player one experiences different trajectories. However, the value function trajectory that player one follows is always bounded between the thresholds we derived from Proposition 3. As Fig. 3 shows, there does not seem to be any direct correlation between player two's discount factor and player one's value function. The interval bounds we derived do tightly approximate resulting value function trajectories.

Alternatively, suppose that both players have the same discount factor but whether player two is minimizing or maximizing its objective is unknown. In Fig. 4, we show player one's value function trajectories for both scenarios: maximizing C^2 (player one's value functions is shown in dotted lines) and minimizing C^2 (player one's value functions is shown in solid lines). The grey region shows the predicted bounds as derived from Proposition 3. Both player one's and player two's initial value functions are randomly initialized as $V^0_{1s,2s} \in [0, 1], \forall\, s \in [9]$.

As Fig. 4 shows, player two's policy changes cause significant shifts in player one's value function trajectories. When player two attempts to minimize its own cost, player one's value function achieves the lower bound as predicted by Proposition 4. This is

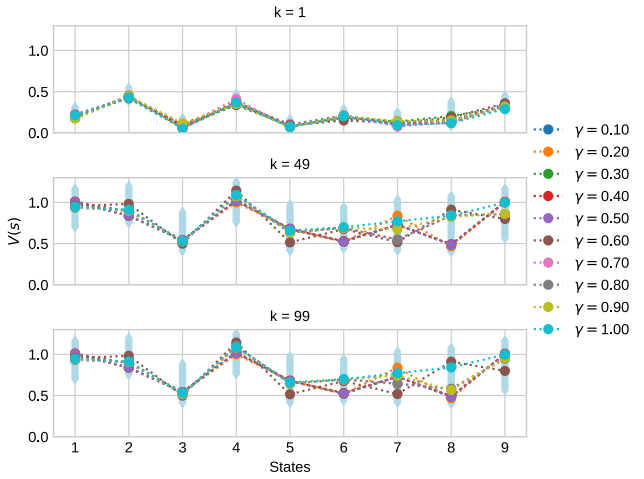


Fig. 3. Player one's value function as a function of the state at different time steps $k = \{1, 49, 99\}$. For each k, the vertical blue lines represent the interval set $\mathcal{V} = [\underline{V}^k, \overline{V}^k]$.

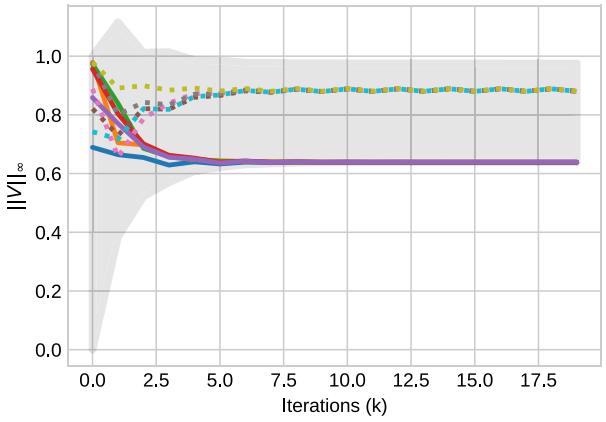


Fig. 4. The infinity norm of player one's value function as a function of iteration k. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

due to the fact that the player objectives do not conflict and at least four actions with different costs are available at each state. Since both players are only selecting from deterministic policies, they are bound to select different actions unless all actions have the exact same cost. On the other hand, if player two is maximizing its value function, then both players would precisely select the same actions from every state. Then depending on the coupling matrix J, they may or may not choose a less costly action at the next step. This causes the limit cycle behaviour that the dotted trajectories exhibit. In terms of the tightness of the bounds we derived in Proposition 3, we note that Fig. 4 also demonstrates the existence of trajectories approaching both the upper and lower bounds of our fixed point set, therefore demonstrating that the set-based bounds are tight in practice.

8. Conclusion

We have bounded the set of optimal value functions for the set-based Bellman operator associated with a discounted infinite horizon MDP. Our results are motivated by parameter-uncertain MDPs and value functions trajectories of a player in stochastic games. We demonstrate our example on a grid MDP and show that while player one's value function does not converge, the Hausdorff distance between the value function and the fixed point set of the set-based Bellman operator converges to zero. Future work includes extending the set-based analysis to consider uncertainty in the transition kernels to fully bounding value function trajectories of learning algorithms in a general stochastic game.

References

Abbad, M., & Filar, J. A. (1992). Perturbation and stability theory for Markov control problems. *IEEE Transactions on Automatic Control*, *37*(9), 1415–1420.

Açıkmeşe, Behçet, & Bayard, David S. (2015). Markov chain approach to probabilistic guidance for swarms of autonomous agents. *Asian Journal of Control*, *17*(4), 1105–1124.

Altman, Eitan (1994). Flow control using the theory of zero sum Markov games. *IEEE Transactions on Automatic Control*, *39*(4), 814–818.

Altman, Eitan, & Gaitsgory, Vladimir A. (1993). Stability and singular perturbations in constrained Markov decision problems. *IEEE Transactions on Automatic Control*, *38*(6), 971–975.

Ang, Samuel, Chan, Hau, Jiang, Albert Xin, & Yeoh, William (2017). Game-theoretic goal recognition models with applications to security domains. In *Int. conf. decision game theory secur.* (pp. 256–272). Springer.

Bellemare, Marc, Dabney, Will, Dadashi, Robert, Taiga, Adrien Ali, Castro, Pablo Samuel, Le Roux, Nicolas, et al. (2019). A geometric perspective on optimal representations for reinforcement learning. In *Adv. neural inf. process. syst.* (pp. 4358–4369).

Bielecki, Tomasz R., & Filar, Jerzy A. (1991). Singularly perturbed Markov control problem: Limiting average cost. *Annals of Operations Research*, *28*(1), 153–168.

Bu, Lucian, Babu, Robert, De Schutter, Bart, et al. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(2), 156–172.

Chatterjee, Krishnendu, Majumdar, Rupak, & Jurdziński, Marcin (2004). On Nash equilibria in stochastic games. In *Int. workshop comput. sci. log.* (pp. 26–40). Springer.

Chavent, Marie (2004). A hausdorff distance between hyper-rectangles for clustering interval data. In *Classification, clustering, and data mining applications* (pp. 333–339). Springer.

Dadashi, Robert, Bellemare, Marc G., Taïga, Adrien Ali, Roux, Nicolas Le, & Schuurmans, Dale (2019). The value function polytope in reinforcement learning. In *Int. conf. machine learning* (pp. 1486–1495).

Delage, Erick, & Mannor, Shie (2010). Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, *58*(1), 203–213.

Demir, Nazlı, Eren, Utku, & Açikmeşe, Behçet (2015). Decentralized probabilistic density control of autonomous swarms with safety. *Autonomous Robots*, *39*(4), 537–554.

Dick, Travis, Gyorgy, Andras, & Szepesvari, Csaba (2014). Online learning in Markov decision processes with changing cost sequences. In *Int. conf. machine learning* (pp. 512–520).

Eisentraut, Julia, Křetínskỳ, Jan, & Rotar, Alexej (2019). Stopping criteria for value and strategy iteration on concurrent stochastic reachability games. arXiv preprint arXiv:1909.08348.

El Chamie, Mahmoud, Yu, Yue, Açikmeşe, Behçet, & Ono, Masahiro (2018). Controlled Markov processes with safety state constraints. *IEEE Transactions on Automatic Control*, *64*(3), 1003–1018.

Eldosouky, Abdel Rahman, Saad, Walid, & Niyato, Dusit (2016). Single controller stochastic games for optimized moving target defense. In *2016 IEEE int. conf commun.* (pp. 1–6). IEEE.

Filar, Jerzy, & Vrieze, Koos (2012). *Competitive Markov decision processes*. Springer Science & Business Media.

Ganzfried, Sam, & Sandholm, Tuomas (2009). Computing equilibria in multiplayer stochastic games of imperfect information. In *21st Int. joint conf. artif. intell.*.

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

Haddad, Serge, & Monmege, Benjamin (2018). Interval iteration algorithm for MDPs and IMDPs. *Theoretical Computer Science*, *735*, 111–131.

Henrikson, Jeff (1999). Completeness and total boundedness of the Hausdorff metric. In *MIT undergraduate J. math.*. Citeseer.

Hu, Junling, & Wellman, Michael P. (2003). Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, *4*(Nov), 1039–1069.

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

Kearns, Michael, Mansour, Yishay, & Singh, Satinder (2000). Fast planning in stochastic games. In *Proc. 16th conf. uncertainty artif. intell.* (pp. 309–316). Morgan Kaufmann Publishers Inc.

Li, Sarah H. Q., Adjé, Assalé, Garoche, Pierre-Loïc, & Açıkmeşe, Behçet (2020). Fixed points of the set-based bellman operator. arXiv preprint arXiv:2001.04535.

Li, Sarah H. Q., Yu, Yue, Calderone, Daniel, Ratliff, Lillian, & Acikmese, Behcet (2019). Tolling for constraint satisfaction in Markov decision process congestion games. In *Amer. control conf.* (pp. 1238–1243). IEEE.

Littman, Michael L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Mach. learn. proc. 1994* (pp. 157–163). Elsevier.

Littman, Michael L. (2001). Value-function reinforcement learning in Markov games. *Cognitive Systems Research*, *2*(1), 55–66.

Moore, Ramon E. (1966). *Interval analysis (vol. 4)*. Prentice-Hall Englewood Cliffs, NJ.

Prasad, H. L., Prashanth, L A, & Bhatnagar, Shalabh (2015). Two-timescale algorithms for learning Nash equilibria in general-sum stochastic games. In *Proc. 2015 int. conf. auton. agents multiagent sys.* (pp. 1371–1379). International Foundation for Autonomous Agents and Multiagent Systems.

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

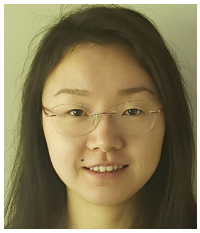
Rudin, Walter, et al. (1964). *Principles of mathematical analysis (vol. 3)*. McGraw-hill New York.

Shapley, Lloyd S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, *39*(10), 1095–1100.

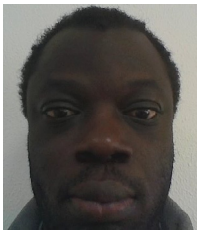
Shiva, Sajjan, Roy, Sankardas, & Dasgupta, Dipankar (2010). Game theory for cyber security. In *Proc. sixth annu. workshop cyber secur. inf. intell. res.* (pp. 1–4).

Wei, Chen-Yu, Hong, Yi-Te, & Lu, Chi-Jen (2017). Online reinforcement learning in stochastic games. In *Adv. neural inf. process. sys.* (pp. 4987–4997).

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



Sarah H.Q. Li received the B.A.Sc degree in Engineering Physics at the University of British Columbia in 2017. Currently, she is working towards her Ph.D degree in Aeronautics and Astronautics Engineering at the University of Washington. 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 multi-agent control.



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 to 2020, he worked at Onera, the French Aerospace Lab. His work is focused on the formal verification of control system software.

Behçet Açıkmeşe is a professor 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 an 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.