The Value Function Polytope in Reinforcement Learning

Robert Dadashi, Adrien Ali Taiga, Nicolas Le Roux, Dale Schuurmans, Marc G. Bellemare

Main Idea

Characterizing the space of value functions in a finite state-action Markov Decision Process context

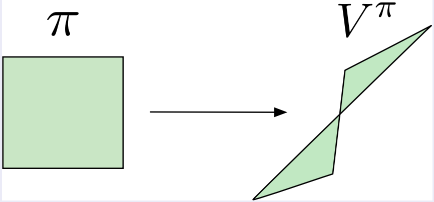
Main Idea

Characterizing the space of value functions in a finite state-action Markov Decision Process context

A geometric perspective

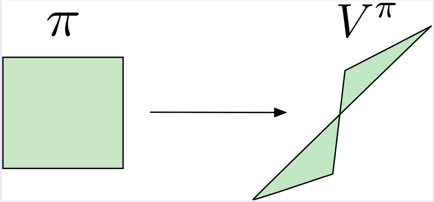
Main Idea

The space of value functions forms a polytope

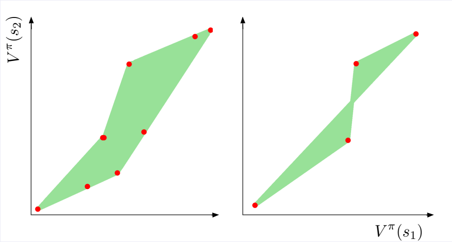


Main Idea

The space of value functions forms a polytope



The boundary of this polytope can be described by value functions corresponding to "semi-deterministic" policies



Preliminaries

We are working with a Markov Decision Process, $M := (\mathcal{S}, \mathcal{A}, r, P, \gamma)$

$$\begin{aligned} \mathcal{S} &: \text{ finite state space} \\ \mathcal{A} &: \text{ finite action space} \\ r &: \text{ reward function} \\ P &: \text{ transition function} \\ \gamma &: \text{ discount factor in } [0,1) \end{aligned}$$

Preliminaries

We are working with a Markov Decision Process, $M := (\mathcal{S}, \mathcal{A}, r, P, \gamma)$

$$\begin{aligned} \mathcal{S} &: \text{ finite state space} \\ \mathcal{A} &: \text{ finite action space} \\ r &: \text{ reward function} \\ P &: \text{ transition function} \\ \gamma &: \text{ discount factor in } [0,1) \end{aligned}$$

π denotes a policy; $\pi : \mathcal{S} \to \Delta\mathcal{A}$. Together with the transition function P, we have state-to-state transition probabilities with respect to policy π:

Preliminaries

We are working with a Markov Decision Process, $M := (\mathcal{S}, \mathcal{A}, r, P, \gamma)$

$$\begin{aligned} \mathcal{S} &: \text{ finite state space} \\ \mathcal{A} &: \text{ finite action space} \\ r &: \text{ reward function} \\ P &: \text{ transition function} \\ \gamma &: \text{ discount factor in } [0,1) \end{aligned}$$

π denotes a policy; $\pi : \mathcal{S} \to \Delta\mathcal{A}$. Together with the transition function P, we have state-to-state transition probabilities with respect to policy π:

$$P^{\pi}(s'|s) := \sum_{a \in \mathcal{A}} \pi(a|s) P(s'|s,a)$$

Preliminaries

The value function at state s is defined as follows:

$$V^{\pi}(s) := \mathbb{E}_{P^{\pi}}\left[\sum_{i=0}^{\infty} \gamma^{i} r(s_i, a_i) | s_0 = s\right]$$

where $r(s, a)$ is the reward function evaluated at state s and action a

Preliminaries

The value function at state s is defined as follows:

$$V^\pi(s) := \mathbb{E}_{P^\pi}\left[\sum_{i=0}^{\infty} \gamma^i r(s_i, a_i) | s_0 = s\right]$$

where $r(s, a)$ is the reward function evaluated at state s and action a

Definition

Let $f_v : \mathcal{P}(\mathcal{A})^\mathcal{S} \to \mathbb{R}^\mathcal{S}$ be the **value functional** mapping the space of policies to their corresponding value functions.

Definitions

Definition

Policy Determinism: A policy π is

- ***s*-deterministic** for $s \in \mathcal{S}$ if $\pi(a|s) \in \{0, 1\}$.
- **semi-deterministic** if it is s-deterministic for at least one $s \in \mathcal{S}$.
- **deterministic** if it is s-deterministic for all states $s \in \mathcal{S}$.

Definitions

Definition

Policy Determinism: A policy π is

- ***s*-deterministic** for $s \in \mathcal{S}$ if $\pi(a|s) \in \{0, 1\}$.
- **semi-deterministic** if it is s-deterministic for at least one $s \in \mathcal{S}$.
- **deterministic** if it is s-deterministic for all states $s \in \mathcal{S}$.

Definition

Let $s_1, \ldots, s_k$ be states and π a policy. Then $\boldsymbol{Y}^{\boldsymbol{\pi}}_{\boldsymbol{s_1,\ldots,s_k}} \subset \mathcal{P}(\mathcal{A})^{\mathcal{S}}$ is the set of policies that agree with π on the states $s_1, \ldots, s_k$. Similarly, let $\boldsymbol{Y}^{\boldsymbol{\pi}}_{\boldsymbol{\mathcal{S}-\{s\}}}$ be the set of policies which agree with π on all states except s.

Visualizations

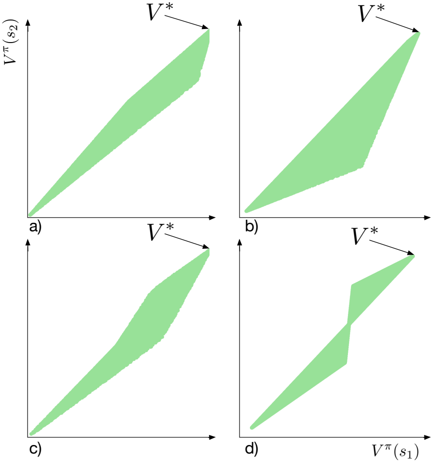


Figure: Value function space corresponding to four two-state MDPs; each evaluated from 50,000 policies sampled uniformly from $\mathcal{P}(\mathcal{A})^{\mathcal{S}}$.

The Line Theorem

Theorem 1: Let s be a state and π a policy. Then there are two s-deterministic policies in $Y^{\pi}_{\mathcal{S}-\{s\}}$, denoted π_l and π_u, such that for all $\pi' \in Y^{\pi}_{\mathcal{S}-\{s\}}$

$$f_v(\pi_l) \preceq f_v(\pi') \preceq f_v(\pi_u)$$

Furthermore, the following are equivalent:

- $f_v(Y^{\pi}_{\mathcal{S}-\{s\}})$
- $\{f_v(\alpha\pi_l + (1-\alpha)\pi_u) : \alpha \in [0,1]\}$
- $\{\alpha f_v(\pi_l) + (1-\alpha) f_v(\pi_u) : \alpha \in [0,1]\}$

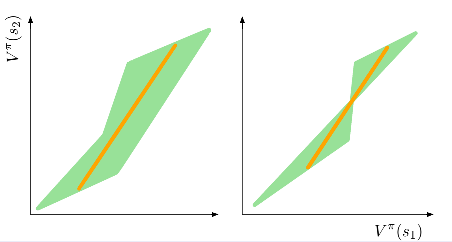
The Line Theorem

Theorem 1: Let s be a state and π a policy. Then there are two s-deterministic policies in $Y^{\pi}_{\mathcal{S}-\{s\}}$, denoted π_l and π_u, such that for all $\pi' \in Y^{\pi}_{\mathcal{S}-\{s\}}$

$$f_v(\pi_l) \preceq f_v(\pi') \preceq f_v(\pi_u)$$

Furthermore, the following are equivalent:

- $f_v(Y^{\pi}_{\mathcal{S}-\{s\}})$
- $\{f_v(\alpha\pi_l + (1-\alpha)\pi_u) : \alpha \in [0,1]\}$
- $\{\alpha f_v(\pi_l) + (1-\alpha) f_v(\pi_u) : \alpha \in [0,1]\}$



Definitions

Definition

A **convex combination** is a finite linear combination of vectors whose coefficients are non-negative and sum to 1.

Definitions

Definition

A **convex combination** is a finite linear combination of vectors whose coefficients are non-negative and sum to 1.

Definition

The **convex hull** of a set E is the intersection of all convex sets containing E. A set C is **convex** if for any $x, y \in C$ $\{\alpha x + (1 - \alpha)y : \alpha \in [0, 1]\} \subset C$.

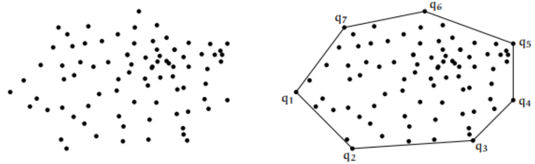
Definitions

Definition

A **convex combination** is a finite linear combination of vectors whose coefficients are non-negative and sum to 1.

Definition

The **convex hull** of a set E is the intersection of all convex sets containing E. A set C is **convex** if for any $x, y \in C$
$\{\alpha x + (1 - \alpha)y : \alpha \in [0, 1]\} \subset C$.



(a) Input. (b) Output.

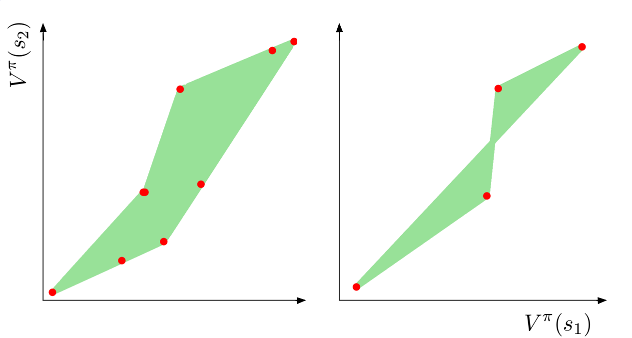
Credit to Harshit Sikchi.

Interesting Consequences of the Line Theorem

For any set of states $s_1, ..., s_k \in \mathcal{S}$ and a policy π, V^π can be expressed as a convex combination of value functions of $\{s_1, ..., s_k\}$-deterministic policies. In particular, $\mathcal{V} := f_v(\mathcal{P}(\mathcal{A})^\mathcal{S})$ is included in the convex hull of the value functions of deterministic policies.

Interesting Consequences of the Line Theorem

For any set of states $s_1, ..., s_k \in \mathcal{S}$ and a policy π, V^{π} can be expressed as a convex combination of value functions of $\{s_1, ..., s_k\}$-deterministic policies. In particular, $\mathcal{V} := f_v(\mathcal{P}(\mathcal{A})^{\mathcal{S}})$ is included in the convex hull of the value functions of deterministic policies.



Interesting Consequences of the Line Theorem

Let V^{π} and $V^{\pi'}$ be two value functions. Then there exists a sequence of policies $\pi_1, ..., \pi_k$ ($k \leq \mathcal{S}$ such that $V^{\pi} = V^{\pi_1}, V^{\pi'} = V^{\pi_k}$, and

$$\{f_v(\alpha\pi_i + (1-\alpha)\pi_{i+1}) : \alpha \in [0,1]\}$$

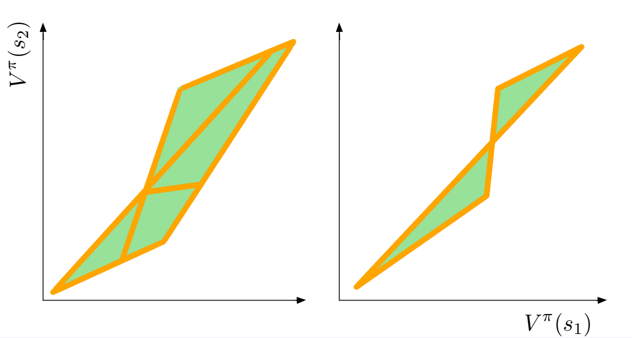
forms a line segment for all $1 \leq i < k$.

Boundary of Semi-Deterministic Policies

The boundary of the space of value functions is a subset of value functions corresponding to semi-deterministic policies.

Boundary of Semi-Deterministic Policies

The boundary of the space of value functions is a subset of value functions corresponding to semi-deterministic policies.



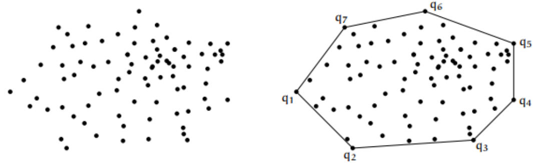
Definitions

Definition

P is a **convex polytope** iff there exist points $x_1, ..., x_k \in \mathbb{R}^n$ such that P is the convex hull of $\{x_1, ..., x_k\}$.

Definition

A **polytope** is a finite union of convex polytopes.



(a) Input. (b) Output.

Credit to Harshit Sikchi.

Main Result

Let π be a policy and let $s_1, ..., s_k$ be states in $\mathcal{S}$. Then $f_v(Y^{\pi}_{s_1,...,s_k})$ is a polytope and in particular, $\mathcal{V} = f_v(Y^{\pi}_{\phi})$ is a polytope.

Main Result

Let π be a policy and let $s_1, \ldots, s_k$ be states in $\mathcal{S}$. Then $f_v(Y^\pi_{s_1,\ldots,s_k})$ is a polytope and in particular, $\mathcal{V} = f_v(Y^\pi_\phi)$ is a polytope.

- Surprising since f_v is in general non-linear and mixtures of policies can describe curves
- There is a sub-polytope structure in the space of value functions