INTERPRETING REINFORCEMENT LEARNING WITH SHAPLEY VALUES

Daniel Beechey, Thomas M. S. Smith, Özgür Şimşek

Bath Reinforcement Learning Laboratory

OVERVIEW

Reinforcement learning provides a rich framework for creating intelligent agents that adapt and improve through continuous interaction with the world. However, uninterpretable agent behaviour hinders the deployment of reinforcement learning at scale.

CONTRIBUTION

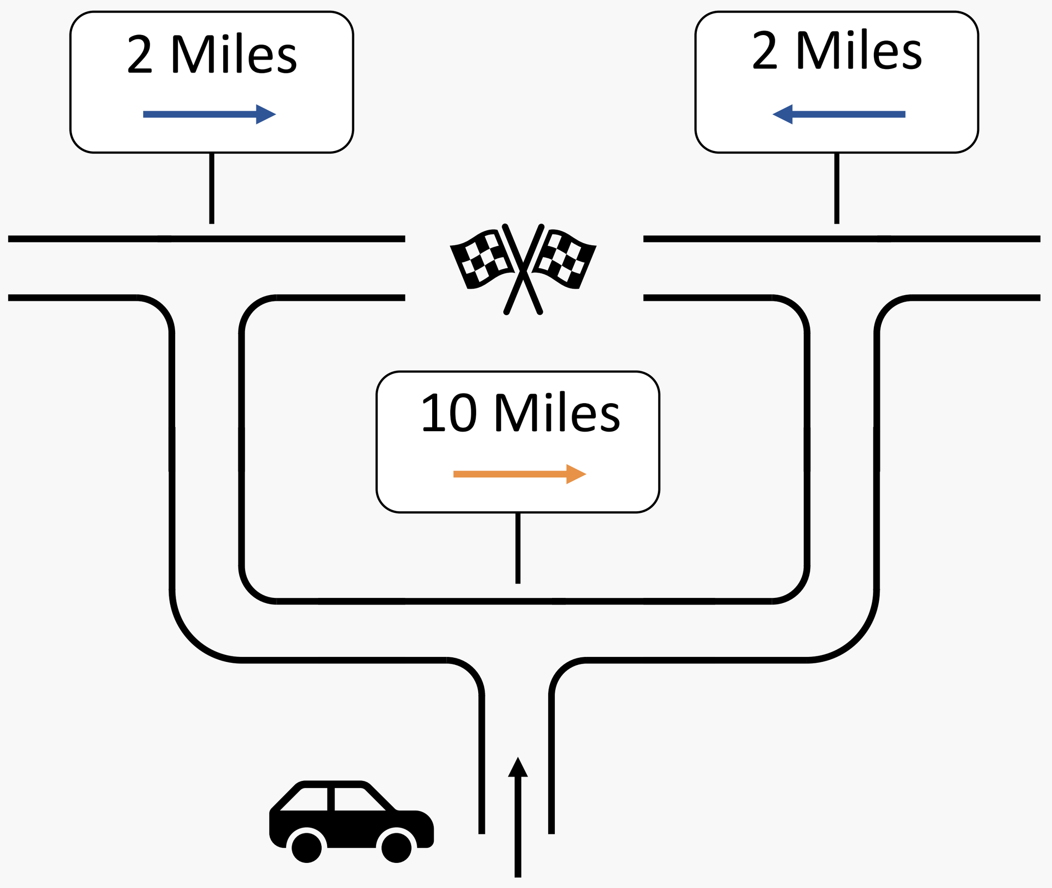
We introduce Shapley Values for Explaining Reinforcement Learning (SVERL), a mathematical framework for explaining agent-environment interactions in reinforcement learning.

In simple domains, SVERL produces meaningful explanations that match human intuition. In complex domains, the explanations reveal novel insight.

WHAT NEEDS EXPLAINING?

Certain features of an agent's observations influence different aspects of agent-environment interactions: **policy, performance and value prediction.**

Example: Autonomous vehicle using signs with directions and distances (features) to navigate the shortest path to a destination.



(a) Directions influence policy.

(b) Directions influence performance (blue arrows) but not always (orange arrow).

(c) Distances influence value prediction but not policy or performance.

COMPUTING FEATURE INFLUENCE

We pose this problem as a contribution assignment problem from cooperative game theory.

A **cooperative game** is a set of players $\mathcal{F}$ and a characteristic function $v\colon 2^{|\mathcal{F}|} \to \mathbb{R}$.

Contribution assignment problem: How to assign the contribution $\phi_i(v)$ of player i to the outcome of the game $(\mathcal{F}, v)$?

Shapley value: $$\phi_i(v) = \sum_{\mathcal{C} \subseteq \mathcal{F} \setminus \{i\}} \frac{|\mathcal{C}|!\,(|\mathcal{F}| - |\mathcal{C}| - 1)!}{|\mathcal{F}|!} [v(\mathcal{C} \cup \{i\}) - v(\mathcal{C})]$$

Shapley values are the unique solution satisfying four axioms specifying the fair distribution of credit across players.

Email: djeb20@bath.ac.uk

Website: djeb20.github.io

SHAPLEY VALUES FOR EXPLAINING REINFORCEMENT LEARNING (SVERL)

Three cooperative games played by the values of features $\mathcal{F}$ at state s whose outcomes are different aspects of agent-environment interactions.

1. EXPLAINING POLICY

Game outcome: $\pi_s^a\colon 2^{|\mathcal{F}|} \to [0, 1]$

The probability of selecting action a at state s when only the values of features $\mathcal{C}$ are known.

$$\pi_s^a(\mathcal{C}) \stackrel{\text{def}}{=} \mathbb{E}[\pi(S, a) \mid S_{\mathcal{C}} = s_{\mathcal{C}}] = \sum_{s' \in \mathcal{S}} p^\pi(s' \mid s_{\mathcal{C}}) \pi(s', a).$$

Shapley values: The contribution of feature values to the probability of selecting action a in state s.

2. EXPLAINING PERFORMANCE

Game outcome: $v_s^\pi\colon 2^{|\mathcal{F}|} \to \mathbb{R}$

The expected return from state s when policy π only knows the values of features $\mathcal{C}$ at state s.

$$v_s^\pi(\mathcal{C}) \stackrel{\text{def}}{=} \mathbb{E}_\mu [G_t \mid S_t = s], \text{where } \mu(s_t, a_t) = \begin{cases} \pi_{s_t}^{a_t}(\mathcal{C}) & \text{if } s_t = s, \\ \pi(s_t, a_t) & \text{otherwise.} \end{cases}$$

Shapley values: The contribution of feature values to expected return from state s.

3. EXPLAINING VALUE PREDICTION

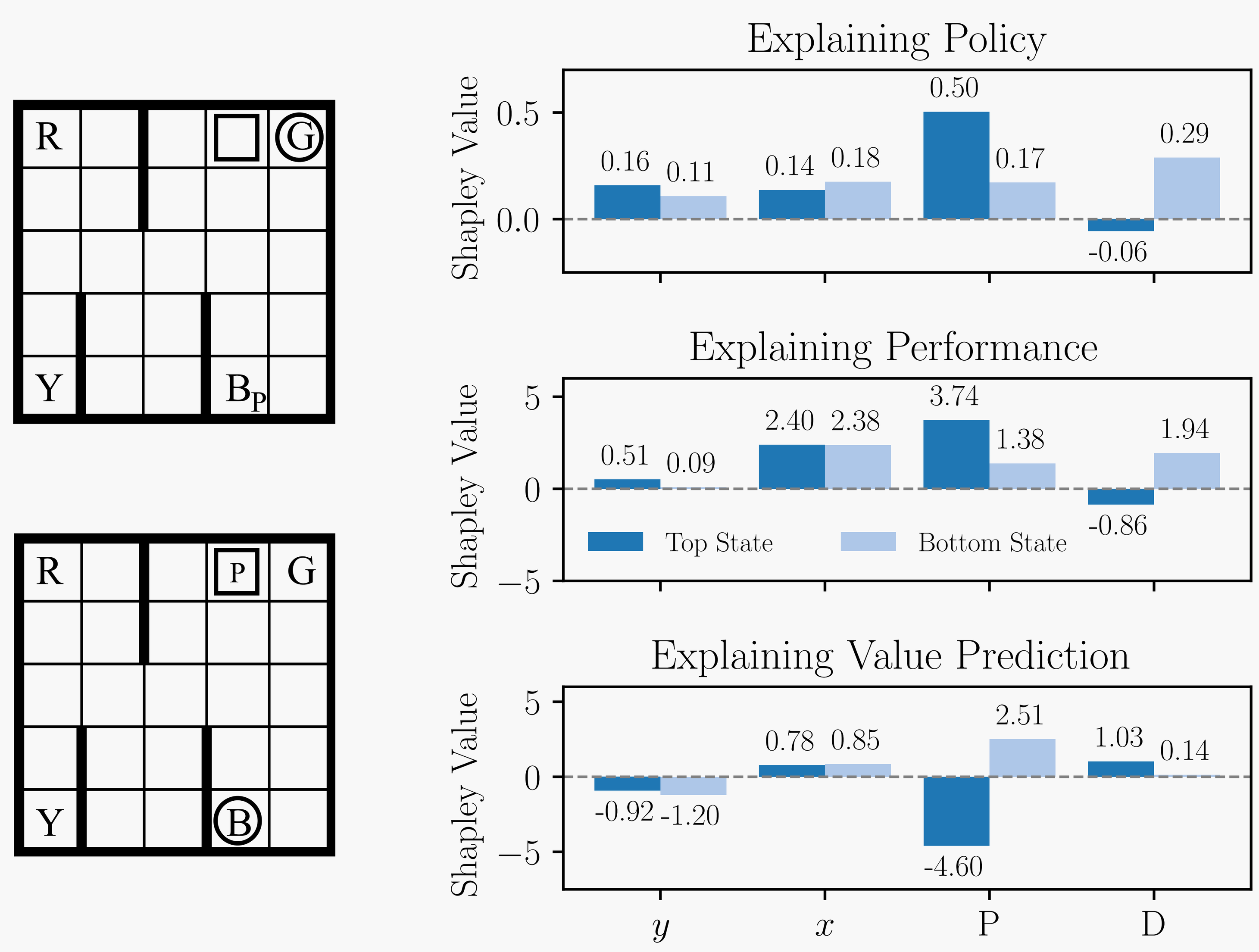
Game outcome: $V_s^\pi\colon 2^{|\mathcal{F}|} \to \mathbb{R}$

The expected return from observation $s_{\mathcal{C}}$ when following policy π.

$$V_s^\pi(\mathcal{C}) \stackrel{\text{def}}{=} \mathbb{E}[v^\pi(S) \mid S_{\mathcal{C}} = s_{\mathcal{C}}] = \sum_{s' \in \mathcal{S}} p^\pi(s' \mid s_{\mathcal{C}}) v^\pi(s').$$

Shapley values: The contribution of feature values to predicting expected return from state s.

EXPLAINING TAXI



Features: Taxi coordinates (x, y), passenger location (P) and destination location (D).

[1] Beechey, D., Smith, T.M. and Şimşek, Ö., 2023, July. Explaining reinforcement learning with Shapley values. In International Conference on Machine Learning (pp. 2003-2014). PMLR.

[2] Lloyd S Shapley. A value for n-person games. 1953.