EXPLAINING REINFORCEMENT LEARNING WITH SHAPLEY VALUES

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

Bath Reinforcement Learning Lab

OVERVIEW

To be widely adopted, it is useful for reinforcement learning systems to not only perform well but also be **explainable**. We introduce **Shapley Values for Explaining Reinforcement Learning (SVERL)**, a theoretical framework for explaining the value predictions, policy and performance of reinforcement learning agents.

SHAPLEY VALUES

Shapley values identify the contributions of individual players to the outcome of a cooperative game. They are the unique solution to a set of mathematical axioms that specify fair distribution of credit across players.

A **cooperative game** is defined by a set of players $\mathcal{F}$ and a characteristic value function $v\colon 2^{|\mathcal{F}|} \to \mathbb{R}$. The Shapley value of player i in game $(\mathcal{F}, v)$ is:

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

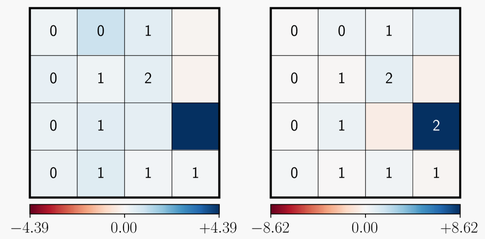
CONTRIBUTIONS

SVERL explains the value function, policy, and performance of an agent using state features.

1. We show that earlier uses of Shapley values in reinforcement learning are incorrect or incomplete.
2. We argue that explaining agent performance is important and overlooked.
3. We develop a principled approach that uses Shapley values to identify the contributions of state features to agent performance.

SVERL produces meaningful explanations that match and supplement human intuition.

Consider the two successive states in Minesweeper below. The features are the 16 grid squares, with possible values 0, 1, 2, or unopened. Each square's shading shows its Shapley value. SVERL clearly shows that one square contributes the most to performance. It is intuitively the most important feature because the locations of both mines can be deduced **only** when this square is opened.



SHAPLEY VALUES FOR EXPLAINING REINFORCEMENT LEARNING (SVERL)

1. EXPLAINING THE VALUE FUNCTION

Characteristic value function: $v^{\hat{V}}(\mathcal{C}) \coloneqq \hat{V}^{\pi}_{\mathcal{C}}(s) = \sum_{s' \in \mathcal{S}} p^{\pi}(s'|s_{\mathcal{C}}) \hat{V}^{\pi}(s')$

Explains the **predictions of the value function** under the assumption that all features will be observed by the agent when acting in the environment.

2. EXPLAINING THE POLICY

Characteristic value function: $v^{\pi}(\mathcal{C}) \coloneqq \pi_{\mathcal{C}}(a|s) = \sum_{s' \in \mathcal{S}} p^{\pi}(s'|s_{\mathcal{C}}) \pi(a|s')$

Explains the **probability of selecting each action**.

3. EXPLAINING PERFORMANCE

1. **Shapley values applied to V^{π}** do not derive the new policy when features are removed and hence cannot show the contributions of features to behaviour or performance.
2. **Shapley values applied to π** show the contributions of features to behaviour but not to agent performance.
3. **SVERL-Performance (SVERL-P)** derives and evaluates the expected return of a new policy when features are removed and hence shows the contributions of features to performance.

Local characteristic value function:

$$v^{\text{local}}(\mathcal{C}) \coloneqq \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s\right]$$

$$\text{where } \hat{\pi}(a_t|s_t) = \begin{cases} \pi_{\mathcal{C}}(a_t|s_t) & \text{if } s_t = s \\ \pi(a_t|s_t) & \text{otherwise} \end{cases}$$

Explains **local agent performance** from state s.

Global characteristic value function:

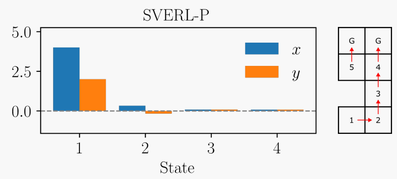
$$v^{\text{global}}(\mathcal{C}) \coloneqq \mathbb{E}_{\pi_{\mathcal{C}}}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s\right]$$

$$\Phi_i(v^{\text{global}}) = \mathbb{E}_{p^{\pi}(s)}\left[\phi_i\left(v^{\text{global}}, s\right)\right]$$

Explains **global agent performance** from all states.

EXAMPLES

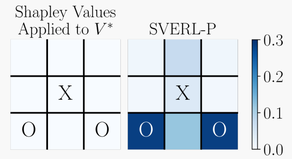
GRIDWORLD



Features are x and y; optimal policy is shown in red.

In states 3 and 4, x or y is sufficient to take the optimal action N so they contribute equally to performance. **In state 2**, observing x is sufficient to choose N while observing only y decreases the probability of choosing N. Overall x contributes positively to performance and y contributes negatively. **In state 1**, x is sufficient to take the optimal action E (because the agent never visits state 5 under the optimal policy) but y is not. Observing y nonetheless increases the probability of choosing E. Hence both x and y positively contribute to performance, with x contributing more.

TIC-TAC-TOE

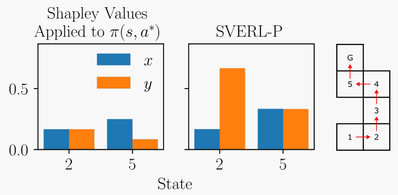


Features are grid squares, taking possible values X, O or empty. The agent plays as X against a Minimax opponent as O.

The two squares marked by the opponent inform the agent that it needs to make a blocking move.

SVERL-P shows that these two squares contribute the most to the agent's performance, aligning with human intuition on the influence of state features. Shapley values applied to V^* on the other hand show that no features contribute to predicting expected return. This is because V^* is always zero in this example, independent of state.

GRIDWORLD



Features are x and y; optimal policy is shown in red.

Shapley values applied to π show that x contributes more to the probability of selecting the optimal action N in state 5. But it would be incorrect to conclude that x is more important than y to act optimally.

SVERL-P shows that x and y contribute equally to performance.

The reason for this difference is that, in state 5, x also contributes towards the likelihood of selecting the worst action (E).