EXPLAINING REINFORCEMENT LEARNING WITH SHAPLEY VALUES

Daniel Beechey, Thomas M. S. Smith, Özgür Şimşek
Bath Reinforcement Learning Lab

OVERVIEW

It is important 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 $v: 2^{|\mathcal{F}|} \to \mathbb{R}$. The Shapley value of player i in game $(\mathcal{F}, v)$ is:

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

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.

SHAPLEY VALUES FOR EXPLAINING REINFORCEMENT LEARNING (SVERL)

1. EXPLAINING THE VALUE FUNCTION (SVERL-V^π)

Characteristic value function:

$$v^{\hat{V}}(\mathcal{C}) := \hat{V}^\pi_{\mathcal{C}}(s) = \sum_{s' \in \mathcal{S}} p^\pi(s'|s_{\mathcal{C}})\hat{V}^\pi(s')$$

Explains the **predictions of the value function.**

2. EXPLAINING THE POLICY (SVERL-π)

Characteristic value function:

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

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

3. EXPLAINING PERFORMANCE (SVERL-P)

1. **SVERL-V^π** does not derive the new policy when features are removed and hence cannot show the contributions of features to behaviour or performance.
2. **SVERL-π** shows the contribution of features to policy but not to agent performance.
3. **SVERL-P** evaluates the expected return of the new policy when features are removed and hence shows the contributions of features to performance.

Local characteristic value function.

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

Explains local agent performance from state s.

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

Global characteristic value Function.

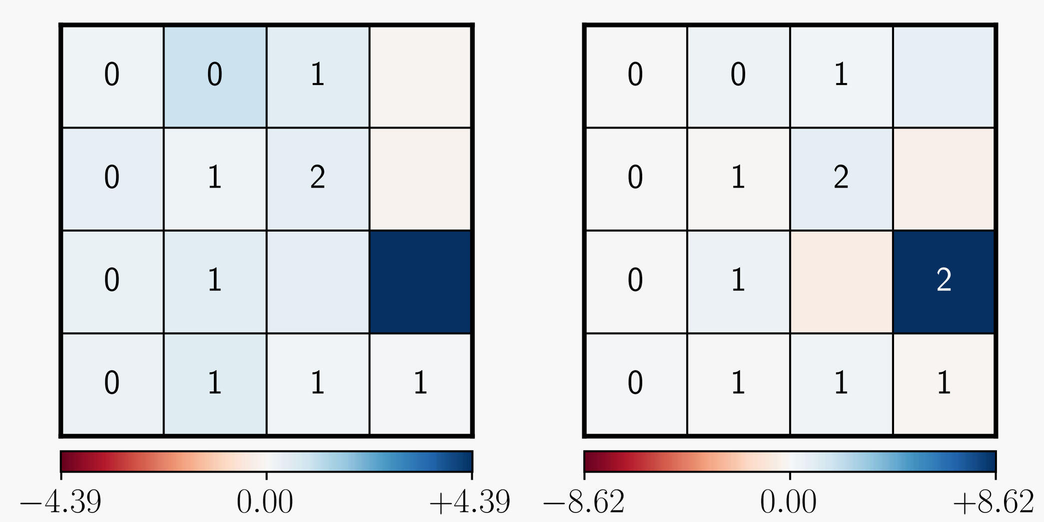
$$v^{\text{global}}(\mathcal{C}) := \mathbb{E}_{\pi_{\mathcal{C}}}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s\right]$$

Explains global agent performance from all states.

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

EXAMPLES

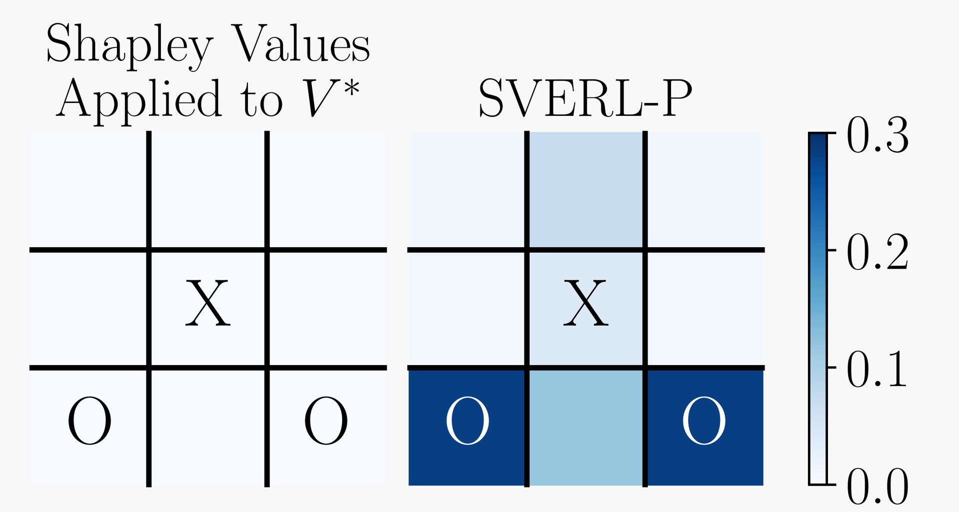
MINESWEEPER



Features: the 16 grid squares, taking possible values 0, 1, 2 or unopened.

SVERL-P shows that one square contributes the most to performance. The locations of both mines can be deduced **only** when this square is opened.

TIC-TAC-TOE



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

SVERL-P shows that two squares contribute the most to performance. SVERL-V^π shows that no features contribute to predicting expected return.