**Towards Interactive Inverse Reinforcement Learning**

**Stuart Armstrong**  **Jan Leike**

---

**Reward learning setup**

The environment $\mu$ is a partially observable Markov decision process, without reward signal. It's $(S, A, O, R, T, \pi_0)$ with $S$ a set of states, $A$ a set of actions, $O$ a set of observations, $T$ a memoryless update rule for states, $O$ a distribution over observations given states, and $\pi_0$ a distribution over initial states $\pi_0$.

A reward function $R : O \rightarrow \mathbb{R}$ is a function that takes a possible observation $o$ and maps it to a numerical reward. The set of possible reward functions is $\mathcal{R}$.

At the end of $m$ steps, the reward-learning algorithm will have learnt which reward functions to value. The reward-learning process is a posterior $P$ which maps $h_m$ (a history of $m$ observations and $n$ actions) to a probability distribution on $\mathcal{R}$.

For shorter histories $h_t$ of length $t < m$, the agent can use a prior $\tilde{P}$ to estimate the distribution.

This posterior gives a value function for a policy $\pi$:

$$V'_\mu(h_t) = \mathbb{E}_P\left[ \sum_{k=t}^m P(h_{k+1} \mid h_k) \sum_{i=k}^m R(h_i) \mid h_t \right]$$

---

**Problem: prevent agent bias and influence**

An agent can value Cooking ($R_1$) or Tidying ($R_2$).

- Typically one reward has higher expected value. An agent that can directly or indirectly choose between $R_1$ and $R_2$ has a biased $P$.
- It may take some time to learn the correct reward. An agent that can randomly decide on one of $R_1$ and $R_2$ has an influenceable $P$.
- If the random decision has the same odds as the learning process, it can be unbiased.

---

**Bias and Influence: example**

An agent can value Cooking ($R_1$) or Tidying ($R_2$).

- Typically one reward has higher expected value. An agent that can directly or indirectly choose between $R_1$ and $R_2$ has a biased $P$.
- It may take some time to learn the correct reward. An agent that can randomly decide on one of $R_1$ and $R_2$ has an influenceable $P$.
- If the random decision has the same odds as the learning process, it can be unbiased.

---

**Bias and Influence: definition**

- $\tilde{P}$ is unbiased for policy $\pi$ if $\mathbb{E}_P[h_t \mid h_t] = \mathbb{E}_P(h_t)$.
- $P$ is unbiased if $\tilde{P}$ is unbiased for all policies $\pi$.
- $P$ is uninformable if there is an observationally equivalent POMDP $R$ in which knowing the initial state $s_0$ determines the distribution: $P(h_t \mid h_0 = s_0)$ is independent of $h_0$.

---

**Bias and Learning**

The agent is uncertain between two reward functions $R_1$ and $R_2$. The horizontal axis shows the current belief $\pi = \mathbb{P}(R_1 \mid h_t)$ over $R_2$, given $h_t$, the history to date. The axis $\pi = 0$ corresponds to certainty in $R_1$, and $\pi = 1$ corresponds to certainty in $R_2$. The function $V'_\mu(h_t)$, the best agent with a specific $\pi$ can get if it never learns more about its reward. It is convex and bounded from above by the convex combination of $V_1''(h_t)$ and $V_2''(h_t)$.

We want to incentivise the agent to perform actions that increase the information about the reward function without increasing bias, i.e., moving up the green arrow. We want to disincentivise the agent to perform actions that manipulate the information about the reward function, i.e., moving along the orange curve (in expectation).

---

**Solutions**

**Bias indifference and bias penalties**

Let $p = \mathbb{P}(R_1 \mid h_t) \in \Delta \mathcal{R}$, where $\mathcal{R}$ is the simplex of probability distributions over $\mathcal{R}$.

For action $a$, let $q(a) = \mathbb{E}_P[h_t \mid h_0, h_t, h_t]$. The action $a$ is biased if $q(a) \neq p$.

Let $R_{\pi(a)} = \sum_{h \in \mathcal{R}} P(h, a, h') R(h')$ and let $V'_{\pi(a)}(h_t) = \mathbb{E}_P[R_{\pi(a)}(h_t) \mid h_t]$, the expected reward from $R_{\pi(a)}$ under the policy $\pi$.

- To make the agent bias-indifferent, the agent will get an extra reward function $R'$, where $R' \leftarrow R + R_{\pi(a)}$.
- To inflict a bias penalty for action $a$, the agent will get an immediate extra reward equal to

$$\max_a \left( V'_\mu(h_t) + V'_{\pi(a)}(h_t) \right) - \max_a \left( V'_\mu(h_t) \right)$$

---

**Counterfactual**

Let $\pi_0$ be any policy, $\mu'$ be any POMDP observationally equivalent to $\mu$ — the same policy results in the same distribution of histories:

$$\mathbb{E}_P[h_t \mid h_0, h_t] = \mathbb{E}_{P'}[h_t \mid h_0]$$

Define $P''_\mu$, the counterfactual posterior:

$$P''_\mu(R \mid h_0) = \sum_{s \in S} P(R \mid s, \pi_0, h_0) \pi_0(s) \mid h_0).$$

Knowing the value of $\pi_0$ fixes $P''_\mu$ entirely, making it uninformable.

There are two natural choices for $\pi_0$:

- A null policy: the agent does nothing. The counterfactual posterior is whatever rewards would have been selected without agent involvement.
- A pure learning policy. The counterfactual posterior is whatever rewards would have been selected with the agent purely learning.