Reward Learning from Multiple Types of Human Feedback

Jason Brown, Carl Henrik Ek, Robert Mullins
University of Cambridge

Reinforcement Learning Pains

Contingent on Reward

Reinforcement learning (RL) is a powerful and general paradigm for applying machine learning to sequential decision-making problems. An agent takes actions in an environment, trying to maximise the output of a reward function which captures the desired behaviour. The agent's policy is optimised based on its past attempts. Despite its generality, **RL is often difficult in practice**, as among many other issues, **good reward functions can be difficult to specify**.

In LLM We Trust?

LLMs might be general purpose reasoners, able to do complex tasks by just asking them. However, they often exhibit **undesirable behaviours** due to their "Shoggoth" nature, which we try and remove via RL from comparative human feedback (RLHF). Additionally, people are trying to apply RL to LLM-based agents. **Can we improve reward specification in these contexts?**

Learning the Reward Function

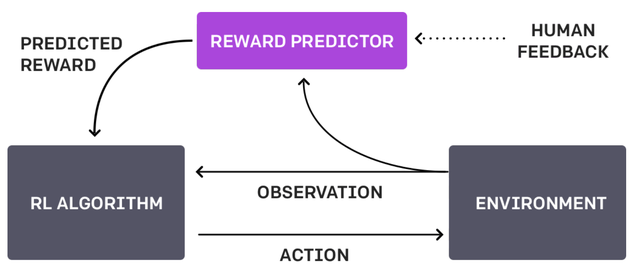

Figure 1: The standard approach

$$P(c^*|R_\theta, C) = \frac{\exp(\beta \cdot \mathbb{E}_{\tau\sim\psi(c^*)}[R_\theta(\tau)])}{\sum_{c\in C}\exp(\beta \cdot \mathbb{E}_{\tau\sim\psi(c)}[R_\theta(\tau)])}$$

Figure 2: Reward-Rational-Choice Theory

Learning from specific feedback types in isolation has been shown to be **effective** across a number of domains. This is typically done via modelling a reward function and updating its parameters via gradient descent (fig. 3). **Inverse RL** (human demonstrations), and **RLHF** (comparative feedback), are the most notable examples of this paradigm.
There is **existing theoretical groundwork laid out for learning from many types of feedback**, framing the human behaviour as a Boltzmann-Rational choice (fig. 4). However, for most practical applications **this formulation is often infeasible**, typically due to infinite continuous choice sets or incomputable expectations.

Human Inspiration

Humans naturally teach using a variety of methods:

- Demonstrating good and bad behaviour
- Providing comparative feedback
- Identifying useful progress measures
- Giving feedback in natural language
- And more...

Utilising many feedback types may be the key for AI to perform complex objectives such as be helpful and harmless or perform intricate physical tasks.

Our Work

Building Solid Practical Foundations

Our key insight is to re-formulate the Reward-Rational-Choice framework (fig. 4) to operate under **restrictive assumptions** that will allow it to **generalise to many domains** in practice. For example, we don't assume a known structure of the trajectory and reward spaces. Here are some highlights of our mathematical foundations.

We perform **MAP estimation** on our reward parameters assuming each feedback type is independent.

$$P(\theta|C, \mathcal{D}) \propto \prod_{C_x\in C} P(C_x|\mathcal{D}, \theta)$$

Reward-rational-choice's deterministic single choice of many is extended, **allowing multiple selections.** There are several ways of doing this, each implying various modelling assumptions as to how the choice was made.

$$P_{\text{RRC}}(c^*|R_\theta, C) = \frac{e^{R_\theta(\psi(c^*))}}{\sum_{c\in C} e^{R_\theta(\psi(c^*))}} \implies$$

$$P_1(C_\mathcal{S}|\mathcal{S}\cup\mathcal{A}, \theta) = \frac{\sum_{\tau\in\mathcal{S}} e^{R_\theta(\tau)}}{\sum_{\tau\in\mathcal{S}} e^{R_\theta(\tau)} + \sum_{\tau\in\mathcal{A}} e^{R_\theta(\tau)}}$$

$$P_2(C_\mathcal{S}|\mathcal{S}\cup\mathcal{A}, \theta) = \frac{e^{p\sum_{\tau\in\mathcal{S}} R_\theta(\tau)}}{e^{p\sum_{\tau\in\mathcal{S}} R_\theta(\tau)} + e^{q\sum_{\tau\in\mathcal{A}} R_\theta(\tau)}}$$

We add an interpretation for **ranked feedback**, and use this for a new interpretation of **demonstrations**.

$$P_{\text{Rank}}(C_r|\mathcal{D}_r, \theta) = \prod_{k=2}^{n} \frac{e^{R_\theta(\tau_k)}}{\sum_{i=1}^{k} e^{R_\theta(\tau_i)}}$$

$$P_{\text{Demo}}(C|\mathcal{D}_a\cup\mathcal{D}_D, \theta) = \prod_{\mathcal{D}_d\in\mathcal{D}_D} P_1(...)P_{\text{Rank}}(...)$$

Preliminary Experimental Results

We have run a variety of experiments across the Mujoco CartPole and HalfCheetah environments, exploring various formulations and combinations of feedback. Here we see how ground truth reward evolves during learning from human feedback when different amounts of preference and demonstrative feedback are available. As expected, **combining feedback types together enables higher performance** than learning on just one of the feedback types alone. More surprisingly, a mixture of feedback types often retains its performance advantage **even when the total amount of human time is kept constant**.

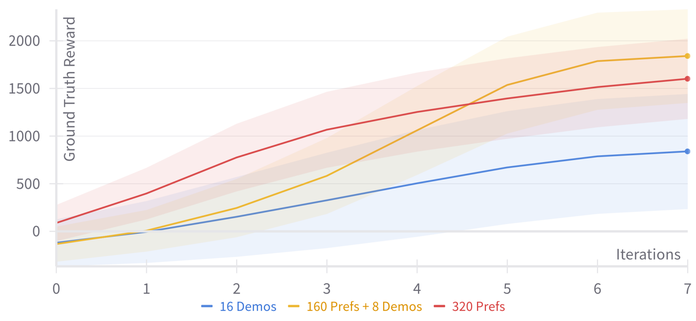

Figure 3: Reward learning on HalfCheetah

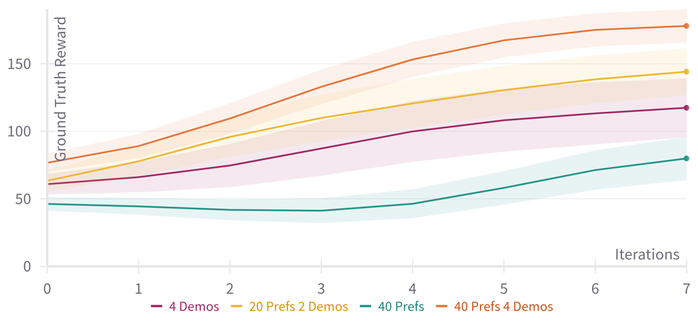

Figure 4: Reward learning on CartPole

Applications and Beyond

Traditional Reinforcement Learning

- Unifies RLHF and Inverse RL
- Allows better learning of tasks with difficult to specify and/or sparse reward functions
- Feedback given by LLMs themselves allows extraction of their internalised knowledge into efficient RL agents
- Greatly expands the possibilities of human-in-the-loop setups
- Achieving more of the vast potential of RL...

LLM Safety & LLM Agents

- Allows utilisation of gold-standard demonstrations for the reward learning step of LLM-RLHF (typically these are just used for supervised fine-tuning of the LLM itself)
- Allows utilisation of many other sources of feedback that may be helpful, such as rankings or multiple proxy reward functions
- Allows distillation of complex LLM agents into smaller expert RL systems, potentially improving efficiency, interpretability, explainability, robustness, and verifiability