Reinforcement Learning in Newcomblike Environments

James Bell[1*] and **Linda Linsefors**[*] and **Caspar Oesterheld**[2*] and **Joar Skalse**[3*]

[1]The Alan Turing Institute, London, UK [2]Duke University, Durham, NC, USA [3]University of Oxford, Oxford, UK
[*]The authors are listed alphabetically.

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

Newcomblike decision problems have been studied extensively in the decision theory literature, but they have so far been largely absent in the reinforcement learning literature. In this paper we study value-based reinforcement learning algorithms in the Newcomblike setting, and answer some of the fundamental theoretical questions about the behaviour of such algorithms in these environments. We show that a value-based reinforcement learning agent cannot converge to a policy that is not *ratifiable*, i.e., does not only choose actions that are optimal given that policy. This gives us a powerful tool for reasoning about the limit behaviour of agents – for example, it lets us show that there are Newcomblike environments in which a reinforcement learning agent cannot converge to any optimal policy. We show that a ratifiable policy always exists in our setting, but that there are cases in which a reinforcement learning agent normally cannot converge to it (and hence cannot converge at all). We also prove several results about the possible limit behaviours of agents in cases where they do not converge to any policy.

Model

Newcomblike decision processes are just like Markov decision processes, except that environment transitions and rewards can depend directly on the agent's policy, see Fig. 1.

Example 1: Newcomb's problem (Nozick, 1969). One state ($S = \{s\}$) (*bandit*), two actions ($A = \{a_1, a_2\}$), reward

$$R(a_1, \pi, s) = \begin{cases} 0 & \text{with probability } \pi(a_2) \\ 10 & \text{with probability } \pi(a_1) \end{cases}$$
$$R(a_2, \pi, s) = \begin{cases} 5 & \text{with probability } \pi(a_2) \\ 15 & \text{with probability } \pi(a_1) \end{cases}$$

Key interesting feature:

- For each policy π, action a_2 does better than a_1.
- But the optimal policy is to take action a_1 with certainty.

Example 2: Death in Damascus (Gibbard and Harper, 1976). One state ($S = \{s\}$) (*bandit*), two actions ($A = \{a_1, a_2\}$), reward

$$R(a_D, \pi) = \begin{cases} 0 & \text{with probability } \pi(a_D) \\ 10 & \text{with probability } \pi(a_A) \end{cases}$$
$$R(a_A, \pi) = \begin{cases} 10 & \text{with probability } \pi(a_D) \\ 0 & \text{with probability } \pi(a_A) \end{cases}$$

Key interesting feature: Each action becomes worse as the agent plays it with higher probability.

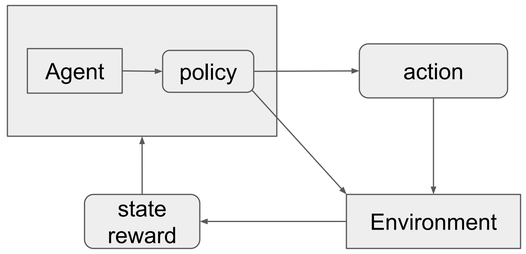

Fig. 1: The distinguishing feature of our model is that the environment can take input directly from the policy.

Model-free reinforcement learning

Research question: How do existing standard learners (specifically, model-free RL agents) behave in Newcomblike decision processes?

Model-free RL:

- Learn values $Q_t(a, s)$.
  Roughly: How well has action a performed empirically in state s?
- Be *greedy in the limit*.
  As $t \to \infty$, choose with high probability action a that approximately maximizes $Q_t(a, s)$.

Model-free RL in Newcomb's problem (Example 1):

- If the agent plays $90\% * a_1 + 10\% * a_2$ and $10\% * a_1 + 90\% * a_2$ equally often, then $Q_t(a_1) > Q_t(a_2)$ with probability 1 as $t \to \infty$.
- If the agent plays any non-deterministic policy forever, then $Q_t(a_2) > Q_t(a_1)$ with probability 1 as $t \to \infty$.

$\Rightarrow$ Assuming sufficient exploration, model-free RL cannot converge to anything other than $100\% * a_2$.

Model-free RL in Death in Damascus (Example 2): Assuming sufficient exploration, model-free RL cannot converge to anything other than $50\% * a_D + 50\% * a_A$ (the optimal policy).

Convergence only to ratifiable policies

Definition (Ratifiability in bandit problems). A policy π is *ratifiable* if

$$\pi(a) > 0 \implies \arg\max_{a \in A} \mathbb{E}\left[R(a, \pi)\right].$$

Ratifiability in our examples: In Newcomb's problem, only $\pi = 100\% * a_2$ is ratifiable. In Death in Damascus, only $\pi = 50\% * a_D + 50\% * a_A$ is ratifiable.

Theorem (Limit policies are ratifiable) (informal). Assume that the agent explores sufficiently, updates Q values appropriately and chooses greedily based on Q. If $\pi_t \to \pi_\infty$ as $t \to \infty$, then π_∞ is ratifiable.

Non-convergence of policies

Theorem 6 (non-convergence of policies) (informal). There are NDPs in which no sufficiently greedy-in-the-limit model-free RL agent can converge to a policy.

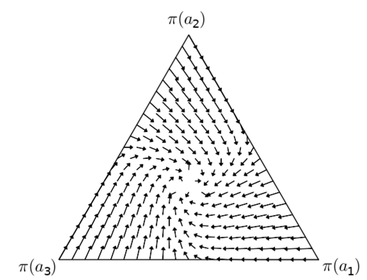

Fig. 2: The softmax learning dynamics of the Repellor problem from our paper, which we use to prove Theorem 6.

References

Gibbard, Allan and William Harper (Feb. 1976). "Counterfactuals and Two Kinds of Expected Utility". In: *Foundations and Applications of Decision Theory* 1, pp. 125–162. DOI: 10.1007/978-94-009-9789-9_5.

Nozick, Robert (1969). "Newcomb's Problem and Two Principles of Choice". In: *Essays in Honor of Carl G. Hempel*. Ed. by Nicholas Rescher et al. Springer, pp. 114–146. DOI: 10.1007/978-94-017-1466-2_7.