Cooperation between Model-Free and Model-Based Strategies in Reinforcement Learning

Aditya Loth	Varun Vankudre	Harsh Agarwal
21111004	20111064	21111030
adityal21@iitk.ac.in	varunsv21@iitk.ac.in	harshmag21@iitk.ac.in

Extended Abstract

There are two primary strategies in the reinforcement learning domain [1]: model-free and model-based. In Model-Free, agent does not utilise environment information; instead, it uses trial and error feedback to update the reward associated with a state-action pair. Thus this model represents habitual behaviour as "an automatic and rapid habitual process linking reward to associated action and enabling the reflexive repetition of previously successful choices." In Model-Based, agent uses environment information to build a decision tree-like structure containing the relationship between state transitions pair and expected reward, making it computationally expensive. This model represents goal-directed behaviour as "a slow, deliberative, goal-directed process comparing the potential outcomes of each action and identifying the action most likely to generate the desired outcome."
Until recently, it was theorised that model-free and model-based compete for dominance over animal behaviour [2], but recent studies have shown that these models can cooperate in some circumstances. In this paper,[3] sequential decision-making experiments are designed and conducted to prove the existence of cooperation between model-free and model-based strategies and to study the working of this hybrid model. The experiment has 4 phases.

- Phase 1(Pre-training phase):- The agent is given two choices, 'A' and 'B' at stage 1, which deterministically transition to stage 2 and stage 3. Instead of feedback in terms of reward, the background and choices of the respective stage are displayed
- Phase 2(training phase):- Reward feedback associated with choices are revealed. Choice 'A' has a greater reward than choice 'B.'
- Phase 3(training phase):- Agent start at either stage 2 or stage 3 with uniform probability. Stage 2 choices 'C' and 'D' are rewarded less than stage 3 choices 'E' and 'F'.
- Phase 4(testing phase): The agent was asked to choose between stage 1 choices 'A' and 'B' without reward feedback but was instructed to make decisions based on previous experience from all three phases.

When tested with humans, they showed retrospective revaluation and chose choice 'B' against choice 'A'. When tested with model-free and model-based agents, they did not show these results. The Dyna architecture replicates human results by integrating model-based planning with model-free learning.

In the previous paper [3], Dyna offers a new way to integrate model-based and model-free. This paper discusses [4] a new approach, Model-Based Pseudoreward Approximation (MBPA), that shapes the reward system to give a computational advantage over Dyna. Pseudorewards imparts a little more information about the reward function to the agent, for an action, which leads to getting closer to goal state gets extra positive reward while action leading to moving away from goal state gets a negative reward. Dyna and MBPA algorithms were executed on two tasks, maze learning and Mountain car problem, and it is observed that MBPA converges faster than Dyna and thus requires less CPU time than Dyna.

References

[1] Y. Huang, Z. A. Yaple, and R. Yu, "Goal-oriented and habitual decisions: Neural signatures of model-based and model-free learning," *NeuroImage*, vol. 215, p. 116834, 2020.

[2] N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan, "Model-based influences on humans' choices and striatal prediction errors," *Neuron*, vol. 69, no. 6, pp. 1204–1215, 2011.

[3] S. J. Gershman, A. B. Markman, and A. R. Otto, "Retrospective revaluation in sequential decision making: a tale of two systems.," *Journal of Experimental Psychology: General*, vol. 143, no. 1, p. 182, 2014.

[4] P. M. Krueger and T. Griffiths, "Shaping model-free habits with model-based goals.," in *CogSci*, 2018.