Information Processing and the Brain CW 2

Kheeran Naidu

December 2019

Introduction

For CW2, I have decided to implement and explore the behaviour of the classical *Reinforcement Learning* algorithm using *Temporal Difference Learning*. Reinforcement learning (RL) was developed from an area of psychology of animal lerning under the name trial-and-error learning [1] and the area of the optimal control problem - using value functions and dynamic programming - in the form of Markov Decision Processes [2][3]. Temporal-difference (TD) methods for RL were proposed as a model of classical (or Pavlovian) conditioning in 1987 [4], and refined to the TD learning rule in 1990 [5]. Put simply, RL is the computational method for goal-oriented learning and the TD learning rule is one which is based on the difference in the current state value and next state value.

Question 1

The RL algorithm is based on the formal framework of Markov decision processes. A Markov decision process is a 4-tuple (S, A, P_a, R_a) that models an environment where S = the set of states, A = the set of actions, $P_a(s, s')$ the transition probability of the action a taking state s to s' and $R_a(s, s')$ the reward function of the action a which takes state s to s'. We often include a discount factor $\gamma \in [0, 1]$ to discount future rewards. In the algorithm, we will exlpoit the Markovian property that the probability of being in a state s_{t+1} with reward $R_a(s_t, s_{t+1})$ at time $t + 1$ is completely described by the state s_t and action a_t at time t, independent of states and actions before time t.

In this CW2 the environment is represented by the 6×8 grid world with walls as obstacles and a single terminal state with reward of 1. In this finite world, the goal is for the agent to obtain the highest reward by following an optimal policy (the actions to take at a particular state) of moving through the environment, which turns out to be the shortest path to the terminal state. Figure **??** represents the grid world where each non-walled tile of the grid is a state and the set of actions are $\{up, down, left, right\}$ which move to the respective adjacent non-walled tiles.

Question 2

The RL algorithm is highly relatable to areas of the brain such as the dopaminergic systems [12] which code for unexpected reward. Recent views discuss

the significant role of the neurotransmitter dopamine in goal-directed behavior, cognition, attention, and reward [10].

Question 3

Question 4

It has been shown that novel sensory stimuli induce similar activity of dopamine cells on unpredicted rewards [9]. The difference being that as the stimuli becomes familiar, the activity diminishes. This has been reasoned by the fact that novelty itself has rewarding characteristics [6] and is intrinsically motivated [8]. The algorithm we have used is a solely extrinsically motivated method of RL, it doesn't cater for the intrinsically motivated rewards that biologcally occurs in animals. Biologically, this intrinsic motivation encourages exploration of an environment which in our algorithm is modelled by the epsilon-greedy mechanism of taking a random action with probability ϵ. However, this approach isn't motivated by the novelty of a stimuli (or state in our case).

Sutton and Barto identify this shortcomming of RL and note that an animal's reward signals are determined by processes within its brain that monitor both the external state and animal's internal state [7]. Barto goes on to develop an *Intrinsically Motivated Reinforcement Learning* model [11]. This model divides the environment into an internal and external environment, as seen in Figure 1. The external enrivonment behaves similarly to our current model, however, the additional internal environment of the model characterises the organism, in particular, it's motivational system. This internal enviroment "*needs to be a sophisticated system that should not have to be redisigned for different problems*", as described by Barto et al.

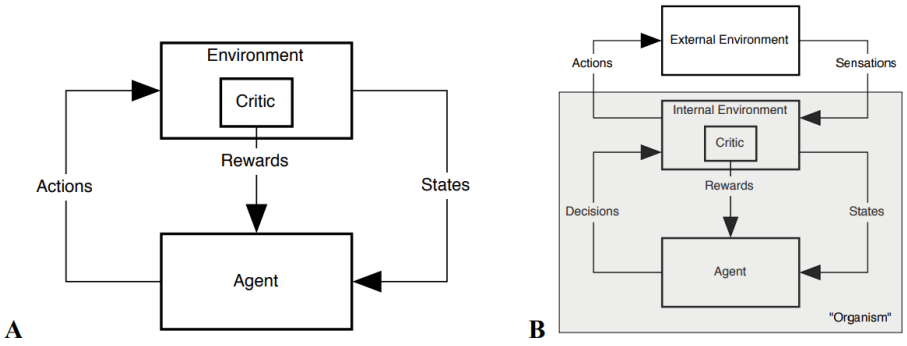

Figure 1: Agent-Environment Interaction in RL. **A**: The usual view. **B**: An elaboration. *Sourced from [11].*

.

References

[1] Robert Sessions Woodworth. "Experimental Psychology. New York: Holt, 1938". In: *Department of Psychology Dartmouth College Hanover, New Hampshire* (1937).

[2] R. E. Bellman. "A markov decision process. journal of Mathematical Mechanics". In: (1957).

[3] R. E. Bellman. "Dynamic programming, princeton univ". In: *Prese Princeton, 1957* (1957).

[4] Richard S Sutton and Andrew G Barto. "A temporal-difference model of classical conditioning". In: *Proceedings of the ninth annual conference of the cognitive science society*. Seattle, WA. 1987, pp. 355–378.

[5] Richard S Sutton and Andrew G Barto. "Time-derivative models of pavlovian reinforcement." In: (1990).

[6] Phil Reed, Chris Mitchell, and Tristan Nokes. "Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task". In: *Animal Learning & Behavior* 24.1 (1996), pp. 38–45.

[7] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*. Vol. 2. 4. MIT press Cambridge, 1998.

[8] Peter Dayan and Bernard W Balleine. "Reward, motivation, and reinforcement learning". In: *Neuron* 36.2 (2002), pp. 285–298.

[9] Sham Kakade and Peter Dayan. "Dopamine: generalization and bonuses". In: *Neural Networks* 15.4-6 (2002), pp. 549–559.

[10] Wolfram Schultz. "Getting formal with dopamine and reward". In: *Neuron* 36.2 (2002), pp. 241–263.

[11] Nuttapong Chentanez, Andrew G Barto, and Satinder P Singh. "Intrinsically motivated reinforcement learning". In: *Advances in neural information processing systems*. 2005, pp. 1281–1288.

[12] Pieter R Roelfsema and Anthony Holtmaat. "Control of synaptic plasticity in deep cortical networks". In: *Nature Reviews Neuroscience* 19.3 (2018), p. 166.