CS698R: Deep Reinforcement Learning Project

Project Report #1

Team Name: DRL_noobs_
Project Title: Foraging in Replenishing Patches
Project #: 7
Team Member Names: Roll No
Parth Srivastava: 190592
Shiven Tripathi: 190816
Archi Gupta: 21111014
Abhinav Joshi: 20211261
Samrudh B Govindaraj: 20128409

1 Introduction

The study of reinforcement learning algorithms plays a vital role in understanding various behavioural effects in learning experiments across species. Where a reward in reinforcement learning has a direct correlation with the results of Dopamine triggers in animals. For learning a task, behavioural predictions derived from the ganglia framework are well studied, including patient, pharmacology, gene studies, etc. (Frank and Fossella [2011], O'Doherty et al. [2004], Samejima et al. [2005], Pessiglione et al. [2006], Lau and Glimcher [2007], Jocham et al. [2011]). The field of reinforcement learning only covers learning a task in an incremental learning setting that relies on prediction errors. In contrast, learning in humans do not exclusively rely on incremental learning and have a more comprehensive array of higher-level functions like working memory (Collins and Frank [2012]). Higher brain activity in the prefrontal cortex is observed during the early stages of learning, making the working memory a significant component for learning a new task.

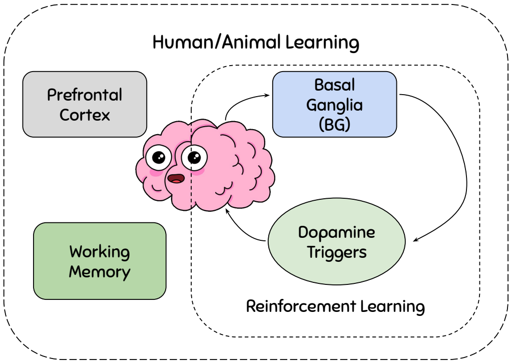

Figure 1: A schematic for Human/Animal Learning process. (missing working memory component in RL designs)

Classical RL models cannot encode behavioural variance present in the learning process of humans. We plan to explore the effect of the working memory module on learning behaviour. We propose an RL model which combines the classical RL methods with a working memory component and studies the effects on Foraging tasks. We plan to use the Foraging with replenishing patches task for our experiments and used its three variants, which require different levels of working memory usage. We further try to understand the role of working memory and sequential decision-making and possibly improve the existing reinforcement learning algorithms.

2 Related Work

Foraging and n-armed bandits and are used to understand the human decision making process (Averbeck [2015]). Certain modeling approaches such as the delta rule reinforcement learning (DDRL) and logistic regression can be used to model decision-making tasks. However, the models assume that the choice of the agent is entirely dependent on the previous outcomes obtained and the future consequences of the decision are not considered. However, MDP-based approaches consider both the previously received outcomes and the future consequences and opportunities of taking a decision. If the agent aims to maximise rewards, MDPs will provide a normative solution.

Foraging tasks have been studied in great detail from a long time. The first progressive work in this direction was the Marginal Value Theorem(MVT), proposed by Charnov [1976]. MVT predicts that the animal should leave the current patch when the energy intake rate within the patch diminishes to the average energy-harvesting rate in the environment.Various experiments have been performed subsequently, testing the viability of this theorem. In one such experiment performed by Constantino and Daw [2015], the subjects had to forage apple trees, deciding when to leave to harvest a tree and to move to a new one. However, the subjects couldn't revisit a tree once it had been visited. The authors compared the returns obtained from the temporal difference algirithms applied to the tasks, and the MVT, and found out that the simple policy derived from MVT ourperforms TD-Lambda Learning. Hall-McMaster and Luyckx [2019] showed that the MVT model better predicts trail-by-error choices in a stay-switch task than a temporal difference RL model. In an other set of interesting experiments, performed by Miller et al. [2017], the authors try to determine the best algorithmic approximation for MVT, using over-wintering waterbowl as a test species. The authors consider three different implementations of the MVT : Classical MVT(MVT), RL based approxiamtion of the MVT (RLMVT) and the Approximation of the MVT using continuously updating estimates(OMVT).). The authors find out that the XRLMVT and the RLMVT perform better than the XOMVT and the OMVT algorithms.

If patches can be revisited, it is possible to model the task as a bandit task. To understand exploration in a two-armed stationary bandit environment, Averbeck [2015] used the Bayes theorem to create a distribution over expected reward probabilities for the two arms. Three factors were found to drive the exploration bonus in the task: the uncertainty in the distribution, the task's time horizon, and the option to switch back to an action. For a non-stationary bandit environment, the relationship between uncertainty and time horizon is seen such that the effect of variance increases if the time horizon is longer. Instrumental learning in humans involves the usage of neural reward circuitry and higher-order mechanisms in prefrontal areas (working memory). Collins and Frank [2012] wanted to see if incorporating higher-order functions such as working memory (WM) in RL models can further explain variance in instrumental learning. Subjects participated in a simple task where they had to learn the correct association between stimulus and response. The number of stimulus-response associations was varied to test the effect of memory on performance. Four computational models were used, a pure RL model, an RL model with forgetting (RLF), an RL+WM model and a pure WM model (algorithm explained in section 5). Actions are chosen probabilistically, incorporating the WM and RL components, weighted by a mixing parameter w(t). The results showed that the RL+WM model provided the best fit for the data across all criteria.

Schonberg et al. [2007] tried to understand the neural basis of the difference in learning and hypothesised that the striatum plays a significant role. A simple four-armed bandit paradigm was used in the task with fixed reward probabilities (0.75-0.25) for each action. An index of how recently actions were chosen was maintained such that if a particular action was chosen, the index value for that action was increased and decayed for the other actions. The probability of selecting an action was calculated by taking the softmax over the sum of the value of the actions and the weighted recency index values. The inverse temperature β and learning rate (a) were used to check reward sensitivity. The results showed that the learners showed a significantly larger sensitivity to rewards (greater $\beta * a$). Learners also had a positive weight for the recency index (perseverence), while non-learners had a negative weight (alternate). Kovach et al. [2012] tried to understand the role of the anterior PFC in the decision making process by comparing performance between patients with brain-lesions and healthy controls. The task contained a four-armed bandit paradigm with rewards for each arm varying in a constrained manner with added Gaussian noise. A conditional logit and a modified TD model was used to model the subject's choices. The conditional logit model calculates the probability of selecting an action by taking the softmax over the product of regression weights and rewards received at the five most recent choices. The modified TD model consists of three independent variables: the actions' value, the recency effect, and the lag difference in actions. The recency effect is similar to the one used by Schonberg et al. [2007]. The lag difference in rewards was calculated as a negative weight for the penultimate reward (t-2) action was observed in the healthy subjects' responses. A softmax over the three variables' weighted sum was used to calculate the probability of selecting a particular action. The results showed that the the frontopolar brain lesion group showed a lower sensitivity to reward trends. The RL+lag difference model was a better fit for the healthy controls choices compared to the pure RL model.

3 Problem Statement

Conventional foraging tasks do not allow the agent to revisit patches. Understanding human foraging behaviour in a replenishing patches environment is an unexplored research area. Building a computational model of human behaviour in the task can shed light on this topic and help understand the role of working memory in sequential decision making.

4 Environment Details and Implementation

We simulate a primary foraging task containing eight bushes present in a unit length octagon for our environment. The agent speed being constant, the distance of unit length corresponds to a unit time in the environment. For each bush, an initial reward is fixed, as shown in Table-1. Further, an agent takes a harvesting action to collect rewards from the bush he stands on. Every harvesting action returns 0.9 of the current rewards present in the bush and replenishes the rewards in other bushes. Replenishing rates of the bushes are fixed for an environment as shown in Table-2. Further, three settings of the environment are fixed as block types. Different Block types in the environment are assumed to have different capacity requirements of working memory. Figure-2 shows the transition of the environment from one state to another.

Different block settings present in our environment would help study the role of working memory in the learning curve of an RL agent. We plan to replicate the human behaviour learning curves with the help of the working memory component added to classical RL frameworks.

Block Type	**Initial Rewards (Time Elapsed = 0)**							
	Bush-1	**Bush-2**	**Bush-3**	**Bush-4**	**Bush-5**	**Bush-6**	**Bush-7**	**Bush-8**
1	0	**70**	**70**	0	**70**	0	**70**	0
2	0	0	**70**	**70**	0	**70**	0	**70**
3	**70**	0	0	**70**	**70**	0	**70**	0

Table 1: Initial Rewards on the Bushes.

Block Type	**Replenishing Rates of Bushes**							
	Bush-1	**Bush-2**	**Bush-3**	**Bush-4**	**Bush-5**	**Bush-6**	**Bush-7**	**Bush-8**
1	0	**4**	**4**	0	**4**	0	**4**	0
2	0	0	**8**	**2**	0	**5**	0	**8**
3	**2**	0	0	**4**	**8**	0	**16**	0

Table 2: Replenishing Rates of Bushes.

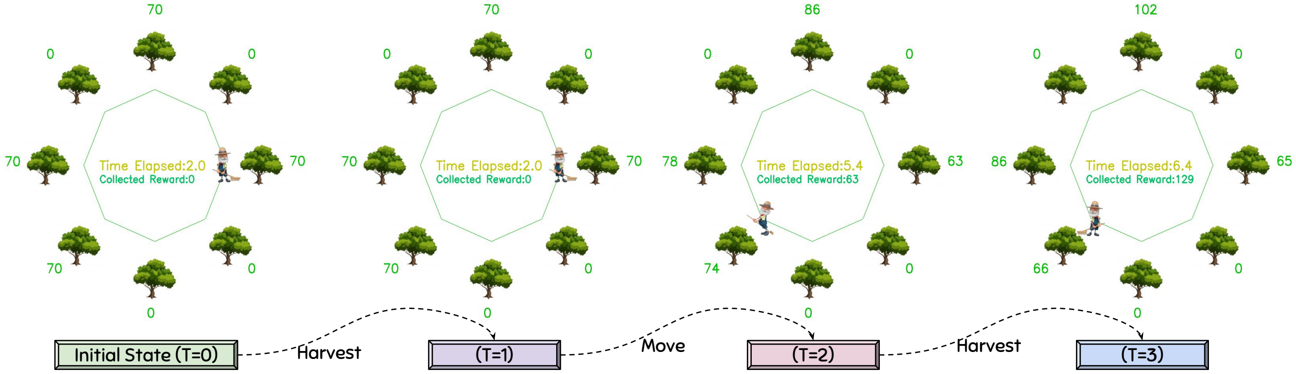

Figure 2: Environment State transition Diagram

5 Future Directions: Prospective Solutions

Firstly, we plan to explore different pure RL online strategies with minor modifications that fit our problem better than multiarmed bandits and to study and compare their performances with each other as well as on different block types. We would be closely observing the final policies in each case and its stability. Simple modifications include smart exploration i.e. after visiting every patch once, the agent explores only rewarding patches, essentially fixing the value function of non rewarding patches in memory as also shown by human behaviour. It should converge faster. Another modification can be having negative rewards as the travelling cost incured by the agent proportional to the distance it travels. We can also experiment with discrete time component in state representation and therefore in value function to better approximate our problem as MDP. Next, we would try blend our model with Marginal Value Theorem by making choice adjustments to a learning rule as suggested by the it (Goldstone and Ashpole [2004]). In foraging tasks, MVT has been observed to give near optimal results and also close to what most animals exhibit.It states that for optimal foraging (in environments with monotonically depleting rewards), it is optimal for the subject to leave the current patch when the reward from that patch drops below the average reward from the environment. Subsequently, we can try to determine the best algorithmic approximation for MVT for our task as given in Miller et al. [2017]. Moreover, it has also been shown that human behaviour in foraging tasks is not well explained by model-free reinforcement learning as humans extrapolate trends in reward rate trajectories to guide foraging choices (Kolling and Akam [2017]. Hence, we also plan to generate optimal policy when the environment is known and thereafter use model based learning.

Finally, we would incorporate human behaviour, particularly Working Memory component in all the above stated models. Mainly, our aim is dynamic integration of RL and WM processes observed in human behaviour to capture behavioural variance. To include a supplementary effect of forgetting across time, an additional step of forgetfulness is introduced after each value funtion updation step. In this additional step, we essentially decay the value function towards their initial values using the folowing equation.

$$Q(s, a) \leftarrow Q(s, a) + \epsilon * (Q_0 - Q(s, a))$$

where $Q_0 = 1/n_A$ is the initial Q value for all actions, representing purely random policy, and controls the degree of forgetfulness. . Thus, more there is the time delay between repeated encounters of a particular state-action, the more its value will be decayed. It basically captures the higher tendency to forget a state if it was encountered earlier. The main algorithm incorporates this forgetfulness of WM by using two value functions : Q_{rl} with pure RL and Q_{wm} with forgetfulness and assigning a weighted probability for action selection. This weight accounts for the limited capacity of WM and varies from individual to individual.

6 Member Contributions

Member	Contributions
Abhinav Joshi (20211261)	Gym Environment, Rendering, Baseline Paper review
Archi Gupta (21111014)	Next Steps Approaches, Baseline Paper review
Parth Srivastava (190592)	Literature Review, Related Works
Samrudh B Govindaraj (20128409)	Literature Reiview, Related Works
Shiven Tripathi (190816)	Environment Testing, Baseline Experimentation

7 Timeline

Project Timeline	Abhinav Joshi	Archi Gupta	Parth Srivastava	Samrudh B Govindaraj	Shiven Tripathi
Sept 30th	Environment Implementation Rendering	Baseline Paper Review	Literature Survey	Literature Survey	Environment Testing
Oct 14th	Report Writing, Baseline paper review	Report Writing	Report Writing	Report Writing, Exp Validation	Report Writing
Oct 28th	RL + WM	RL + WM (exp)	RL + WM (exp)	RL + WM (test)	RL + WM (impl)
Nov 11th	Comparison	Conclusion	Experimentation	Experimentation	Consolidation

References

Bruno B. Averbeck. Theory of choice in bandit, information sampling and foraging tasks. *PLOS Computational Biology*, 11(3), 2015. doi: 10.1371/journal.pcbi.1004164.

Eric L. Charnov. Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, 9(2):129–136, 1976. doi: 10.1016/0040-5809(76)90040-x.

Anne GE Collins and Michael J Frank. How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7): 1024–1035, 2012.

Sara M Constantino and Nathaniel D Daw. Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, 15(4):837–853, 2015.

Michael Frank and John Fossella. Neurogenetics and pharmacology of learning, motivation, and cognition. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 36:133–52, 01 2011. doi: 10.1038/npp.2010.96.

Robert L. Goldstone and Benjamin C. Ashpole. Human foraging behavior in a virtual environment. *Psychonomic Bulletin Review*, 11(3):508–514, 2004. doi: 10.3758/bf03196603.

Sam Hall-McMaster and Fabrice Luyckx. Revisiting foraging approaches in neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, 19(2):225–230, 2019.

Gerhard Jocham, Tilmann A Klein, and Markus Ullsperger. Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *Journal of Neuroscience*, 31(5): 1606–1613, 2011.

Nils Kolling and Thomas Akam. (reinforcement?) learning to forage optimally. *Current opinion in neurobiology*, 46:162–169, 2017.

C. K. Kovach, N. D. Daw, D. Rudrauf, D. Tranel, J. P. Odoherty, and R. Adolphs. Anterior prefrontal cortex contributes to action selection through tracking of recent reward trends. *Journal of Neuroscience*, 32(25):8434–8442, 2012. doi: 10.1523/jneurosci.5468-11.2012.

Brian Lau and Paul W Glimcher. Action and outcome encoding in the primate caudate nucleus. *Journal of Neuroscience*, 27(52):14502–14514, 2007.

Matt L Miller, Kevin M Ringelman, John M Eadie, and Jeffrey C Schank. Time to fly: A comparison of marginal value theorem approximations in an agent-based model of foraging waterfowl. *Ecological Modelling*, 351:77–86, 2017.

John O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science*, 304(5669):452–454, 2004.

Mathias Pessiglione, Ben Seymour, Guillaume Flandin, Raymond J Dolan, and Chris D Frith. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042–1045, 2006.

Kazuyuki Samejima, Yasumasa Ueda, Kenji Doya, and Minoru Kimura. Representation of action-specific reward values in the striatum. *Science*, 310(5752):1337–1340, 2005.

T. Schonberg, N. D. Daw, D. Joel, and J. P. Odoherty. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, 27(47): 12860–12867, 2007. doi: 10.1523/jneurosci.2496-07.2007.