CS698R: Deep Reinforcement Learning Project

Final Project Report

Team Name: DRL_noobs_
Project Title: Foraging in Replenishing Patches
Project #: 7
Team Member Names: Roll No
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.

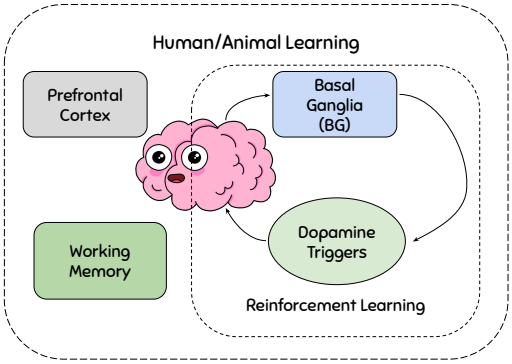

Figure 1: A schematic for Human/Animal Learning process. (missing working memory component in RL designs)

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 to study its effects on Foraging tasks. We use three variants of foraging with replenishing patches task for our experiments and show the performance of various RL algorithms on them. We further propose a mathematical model that combines the Working Memory component with the RL framework and explain its plausibility in detail. Moreover, we also do a detailed ablation analysis of the proposed mathematical model describing each module's impacts on performance. All of our experiments can be found on (https://github.com/iabhinavjoshi/ForagingReplenishingPatches)

2 Related Work

The role of working memory in human learning rates is a well-grown area of research in human psychology. However, all reinforcement learning frameworks lack the property of working memory in an agent learning setting, and the effect of the working memory component on an RL agent is explored less in the reinforcement learning community. Foraging and n-armed bandits and are used to understand the human decision-making process (Averbeck [2015]). Specific modelling approaches such as delta rule reinforcement learning (DDRL) and logistic regression are used to model decision-making tasks. However, the models assume that the choice of the agent is entirely dependent on the previous outcomes obtained, where future consequences of the decision are not considered. However, MDP-based

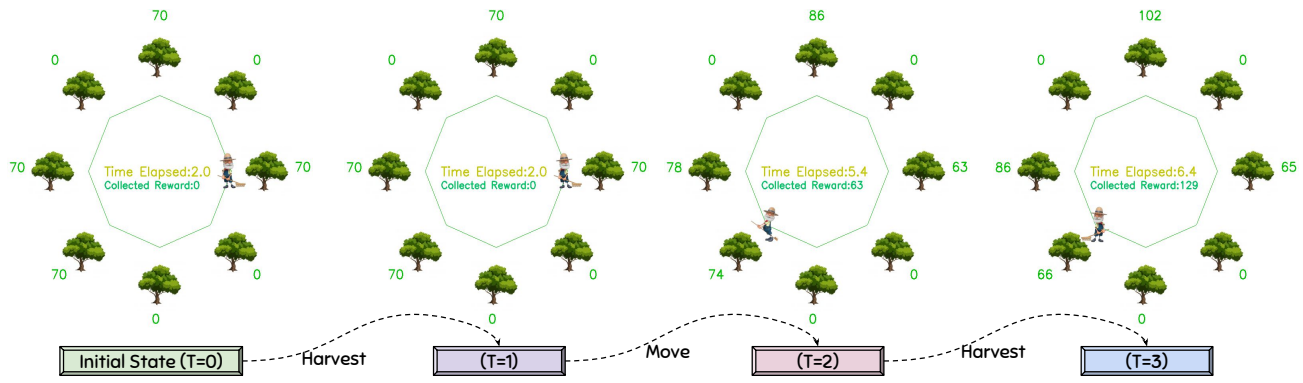

Figure 2: Environment State transition Diagram

approaches consider both the previously received outcomes and the future implications and opportunities of taking a decision. If the agent aims to maximise rewards, MDPs will provide a normative solution.

Foraging tasks has been an active area of research for a long time. Few of the initial and influential works include the Marginal Value Theorem(MVT) Charnov [1976] that predicts that the animal leaving strategy for a patch using the current energy intake rate within the patch and the average energy-harvesting rate in the environment. 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. Another interesting work Miller et al. [2017], perform a comparison between three different implementations of the MVT. Averbeck [2015] studies exploration in a two-armed stationary bandit environment, and use Bayes theorem to create a distribution over expected reward probabilities for the two arms. Schonberg et al. [2007] tries to capture the neural basis of the difference in learning and hypothesise the role of the striatum using a simple four-armed bandit paradigm. Kovach et al. [2012] explores the role of the anterior PFC in the decision making process by comparing performance between patients with brain-lesions and healthy controls using a four-armed bandit paradigm where rewards for each arm vary in a constrained manner with added Gaussian noise. Few of the recent works include, Collins and Frank [2012] which incorporates higher-order functions such as working memory (WM) in RL models and study the variance in instrumental learning.

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. For the simulated foraging task, given a constant time duration t, we have to maximise the reward obtained from harvesting the berries. Breaking this to a more granular level, we have to decide if, at each instant of time t, whether we choose to leave the current patch and explore or we decide to harvest the current patch, subject to the condition that the reward should be maximum.

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 help study the role of working memory in the learning curve of an RL agent. Through extensive experimentation with

RL models, we show a comparison between various RL based approaches and propose a suitable mathematical model for representing human behaviour in such environments.

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 1: Replenishing Rates of Bushes.

5 Proposed Solution

The final solution involves a dynamic integration of RL and WM processes observed in human behaviour to capture behavioural variance Using two value functions Q_{RL} (pure RL) and Q_{WM} (with forgetting). Q_{RL} (pure RL) gives the optimum superhuman solution which is impossible to achieve for a human, whereas Q_{WM} (with forgetting) has a subhuman degrading performance because of forgetful behaviour. Finally to mimic human behaviour we assign a weighted probability for action selection to each of the Q functions. (Collins and Frank [2012])

We do a detailed analysis for marginal value theorem and use it to describe the behaviour of an optimally foraging individual. We identify the problems with MVT in the current context, which include the assumption of no revisiting patches and not considering any storage requirement for previous patches. Moreover, we observe MVT works for decaying rewards but assumes no replenishment, as the concept of revisiting patches is not present.

To overcome these drawbacks, we modify the marginal value theorem and study the obtained theorem in detail. We consider the local rewards for each patch, taking replenishing into account. We further add the idea of estimated rewards and include the memory component in our module.

$$P(\text{ exploit })_t = \frac{1}{1 + \exp\left(-\left[c + \beta\left(r_{t-1} - T_t\right)\right]\right)}$$

6 Experiments

To study the Foraging environment in detail, we perform extensive experiments by making various assumptions and posing the problem in multiple settings. In this section, we overview the settings tried for studying the performance of existing RL algorithms.

6.1 Marginal Value Theorem:

We tried variations of the standard MVT using replenishment rates, local and global thresholds, working memory, action selection strategies and strategies to learn the initial rewards (see Appendix 9). The MVT model along with replenishment rate provided the highest rewards out of all the variants, across the three blocks. Adding replenishment rate to the model provided the highest benefit in block 3, as the difference in replenishment rates is larger in block 3. However, the total rewards reduced for higher proportions of replenishment rates, which can be explained through the higher time spent travelling by the agent Figure 3.The probabilities of staying at the patches and the average thresholds for each patch closely resemble the replenishment rates for the patches. This indicates that the algorithm has obtained a sufficient understanding of the environment as it spends more time at the highly rewarding patches (Figure 4).

6.2 Multi-Armed Bandit and MDP:

For testing the performance of basic RL agents in our environment, we take Foraging task as a multi-armed bandits situation where the agent picks a patch to harvest. Each episode is a sequence of bandit decisions, until time

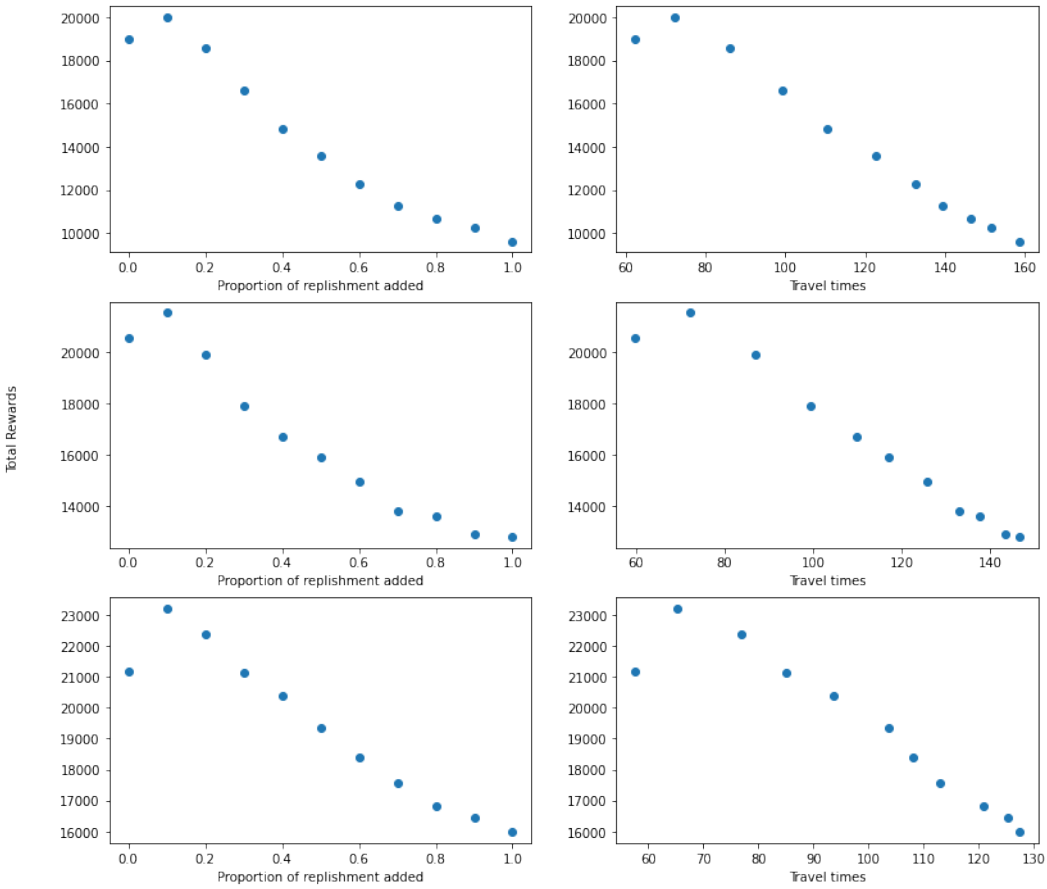

Figure 3: Rewards for replenishment rates and travel times across blocks in MVT

runs out Under these simplifying constraints, we consider strategies which include Pure Greedy, Pure Exploratory, Fixed Epsilon Greedy, Decaying Epsilon Greedy and Uncertainty Confidence Bound (UCB). Figure - 15 shows a comparison between the tried approaches.
The problem can also be approximated as a MDP which can be tested on sarsa, qlearning and sarsa with eligibility traces. Figure 19 and 20 shows their performance on block 1 and 3 respectively.

6.3 Deep RL methods:

In this setting, we treat the foraging task as a Markov Decision Process, where the agent decides to pick a patch for harvesting. Each episode consist of a sequence of decisions until the time runs out We further use Multi-Layer Perceptrons as function approximators for policy and state values.
DQN: Mnih et al. [2013] Q-learning aims to approximate the optimal action-value function: Q (s, a) Q(s, a;) where Q(s, a;) is an approximate action-value function, and represents the learnable parameters. We use an MLP with one hidden layer of 64 dimensions.
A2C: Mnih et al. [2016] A2C is an actor-critic algorithm consisting of two networks (the actor and the critic). The actor-network is used to choose an action at each time step, and the critic network evaluates the estimated Q-values for a given state. As the algorithm learns, the actor-network uses the feedback provided by the critic and learns to improve on a given task. For A2C, we use a similar setting where the function approximator is an MLP with one hidden layer of 64 dimensions and use shared weights for Policy and Value Networks.
Proximal Policy Optimization (PPO): Schulman et al. [2017] This algorithm introduces a clipped surrogate objective over the Actor-Critic Methods

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t \right) \right]$$

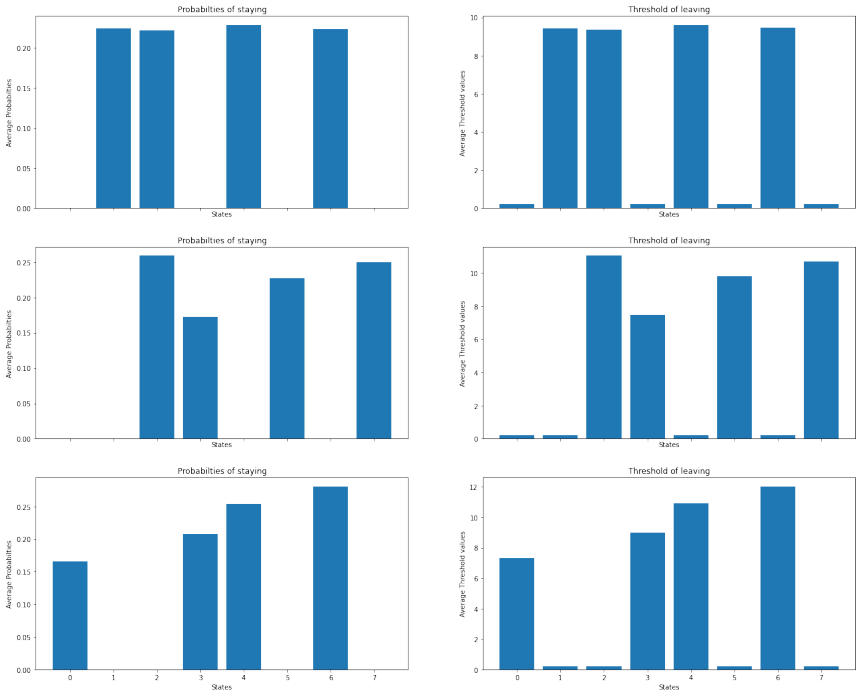

Figure 4: Probabilities of staying and average threshold values for every patch in MVT

In Figure 17, we show a comparison between all the tried deep learning based RL approaches.

6.4 Human Forgetful Behaviour Modelling

We consider the study of human behaviour in foraging tasks as our primary objective. Working memory plays an essential role in human learning. For replicating human behaviour, we propose the integration of RL and the working memory component. First we introduce a element of forgetfulness: A forgetful agent that decays the value function towards their initial values after each value update step.

$$Q(s,a) \leftarrow Q(s,a) + \varepsilon\left(Q_0 - Q(s,a)\right)$$

The more the time delay between repeated encounters of a particular state, action, the more its value will be decayed. We can also decrease the decaying rate with episodes since as we play repeatedly, we become better and gradually we forget less. The final model incorporates working model by using two value functions Q_{RL} (pure RL) and Q_{WM} (with forgetting) and assigning a weighted probability for action selection. The probability is given my c/n_s

$$Q_{RL} \text{ if } \frac{c}{n_s} > 0.5$$

$$Q_{WM} \text{ otherwise}$$

where n_s is the task specific memory load and c is individual memory capacity. We have shown the performance of forgetful RL and WM RL in figure 21 and 22.

7 Results and Analysis

Looking at the performance of the algorithms, Q-Learning seems to converge to the highest rewards in Block 3 and the MVT with replenishment model, performs the best in Block 1. The travel times seem to mostly explain the difference in performance in the algorithms (Figure 7).

Method	Reward in Block 1	Reward in Block 3
MVT (with replenishment)	19983.5	23215.8
Multi Armed Bandit (Decaying Epsilon Greedy)	17199	22102.2
SARSA	15842	20999
Q Learning	18814.7	25732
Deep RL (PPO)	11610.9	16104

Table 2: Rewards Obtained by Trained Agents

Agent	Convergence reward	Avg Harvest Time	Avg Travel Time
sarsa	15842	196.35	103.35
q_learning	18814	183	117
Sarsa lambda	14997	202	98.
Forgetful sarsa	1143	270.8	29.2
forgetdecay sarsa	14624	200.45	99.5
WM_sarsa	10594	140.765	159.235
MVT	19983.5	227.81	72.19

Table 3: Quantitative Evaluation Block-1

Agent	Convergence reward	Avg Harvest Time	Avg Travel Time
sarsa	20999	205.9	94.1
q_learning	25732	176.35	123.65
Sarsa lambda	18339	218.3	81.69
Forgetful sarsa	1591	282.9	15.05
forgetdecay sarsa	19071	207.55	92.45
WM_sarsa ()	13892.355	146.06	153.94
MVT	23215.8	234.83	65.16

Table 4: Quantitative Evaluation Block-3

8 Future Directions and Conclusions

The working memory model has a drawback such that the two value functions, Pure RL and Pure Decay, provide estimated Q values which lie on the extremes of the performance spectrum. In the future, we could try decaying one of the Q functions randomly, such that the estimated Q functions are moderate in comparison. To prevent the Q_{WM} values from decaying very strongly, which causes a large drop in performance, we can push the Q_{WM} estimates to the Pure RL values at regular intervals. This can be supported by the fact that forgetting need not be compounded, and sudden instances of remembering things which have been forgotten have been reported in humans. Memory models can also incorporate the primacy and recency effect to model human behaviour.

9 Member Contributions

Member	Contributions
Abhinav Joshi (20211261)	Gym Environment, Rendering, Literature review, Presentation
Archi Gupta (21111014)	Working Memory, MDP Methods, Literature review, Presentation
Samrudh B Govindaraj (20128409)	Literature Review, MVT and augmentations, Presentation
Shiven Tripathi (190816)	Environment Testing, MAB Methods, Deep RL Methods, Presentation

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.

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.

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.

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.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning, 2016.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Appendix

A Experimental Details

MVT hyperparameter values
Start reward = 50., Local learning rate = 0.99, K = 3 (min values to learn local reward estimates), Noise Std = 1, Initial global reward = 50, Global learning rate = 0.9, Episodes = 100, C = 0.8, Beta = 0.8

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 5: Initial Rewards on the Bushes.

B Extra Plots

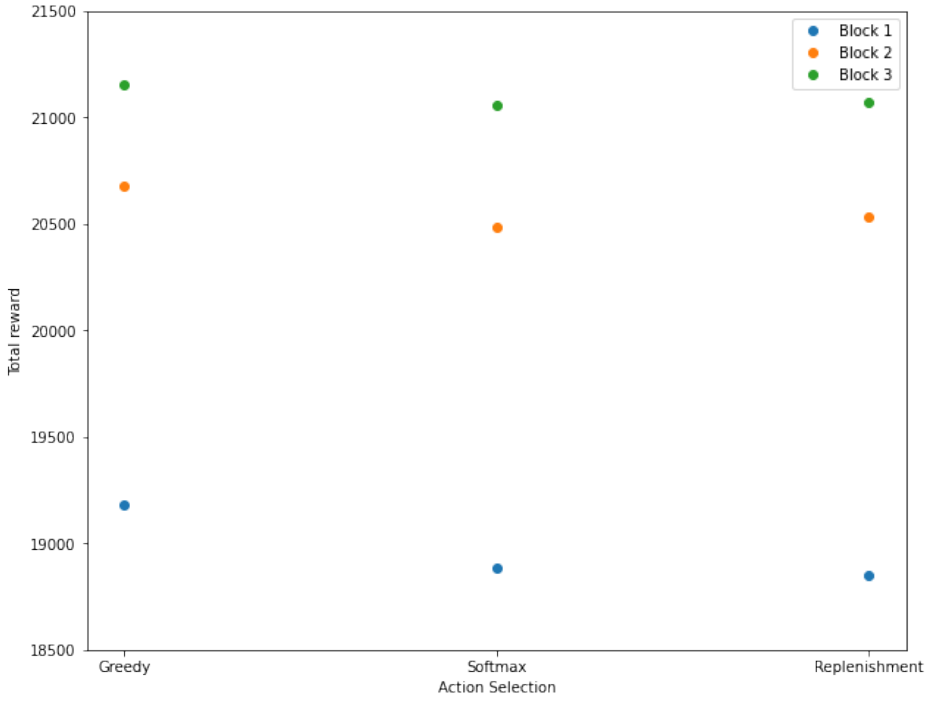

Figure 5: Rewards for action selection strategy across blocks in MVT

C More Information, etc.

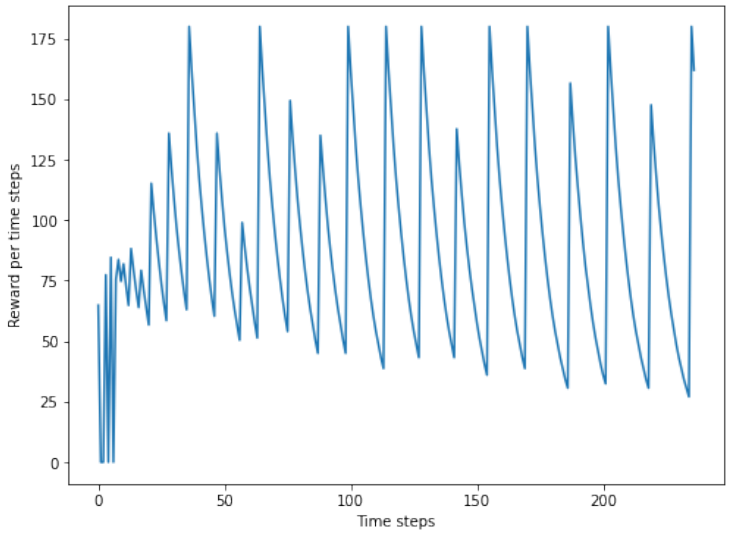

Figure 6: Rewards for reward across time steps in MVT

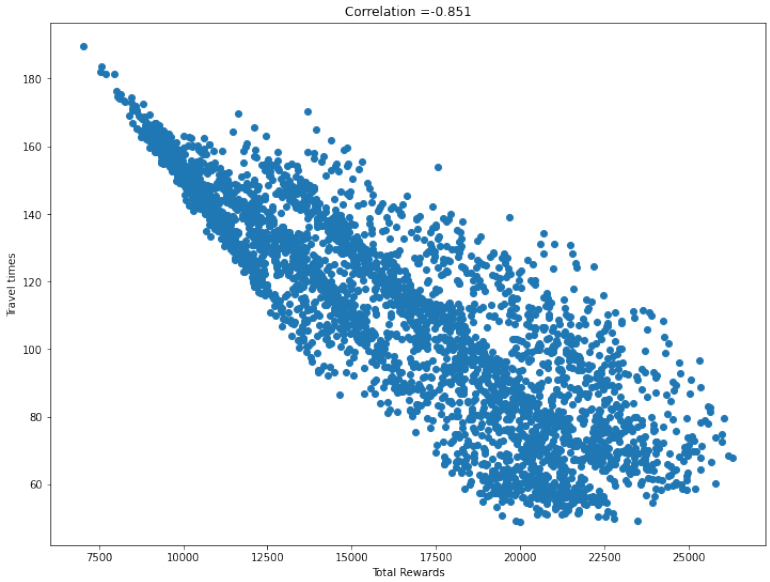

Figure 7: Correlation between rewards and travel time in MVT

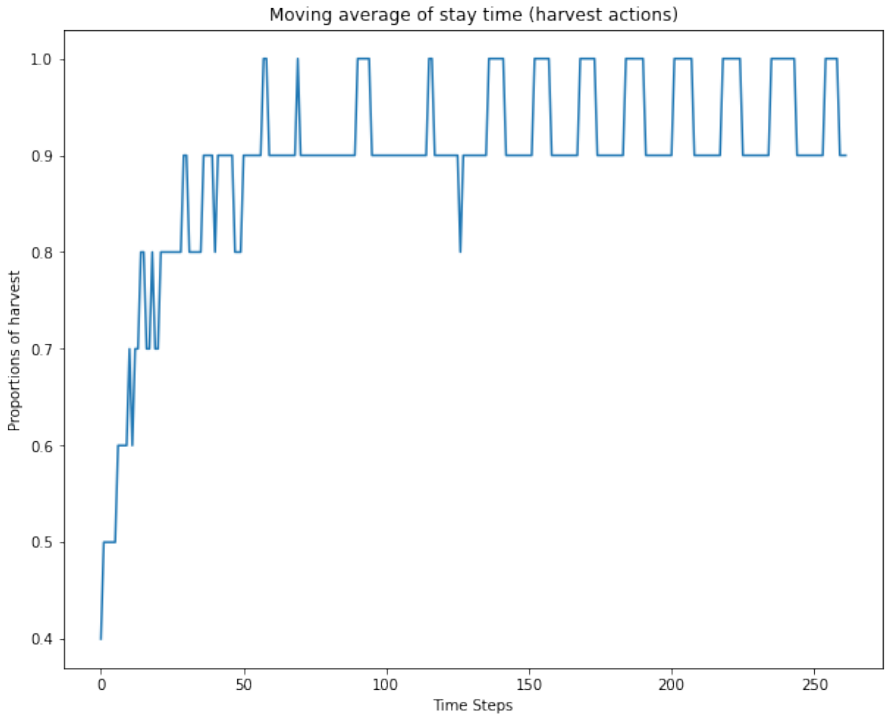

Figure 8: Moving average of harvest actions across time in MVT

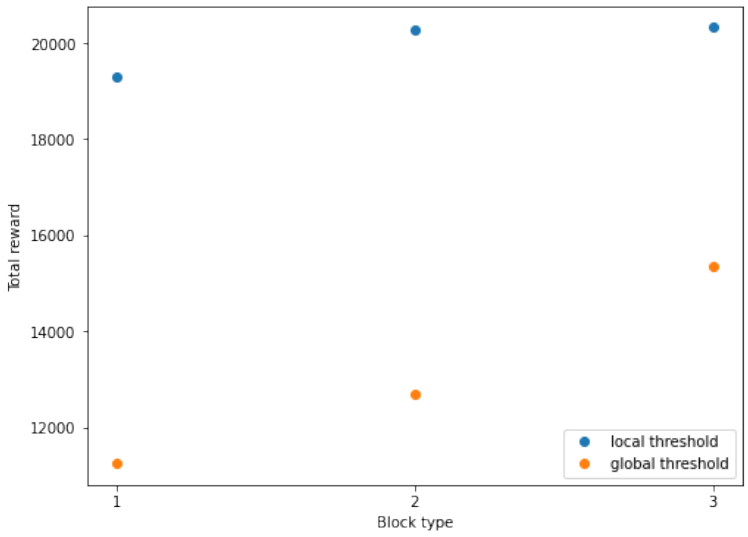

Figure 9: Rewards for threshold types across blocks in MVT

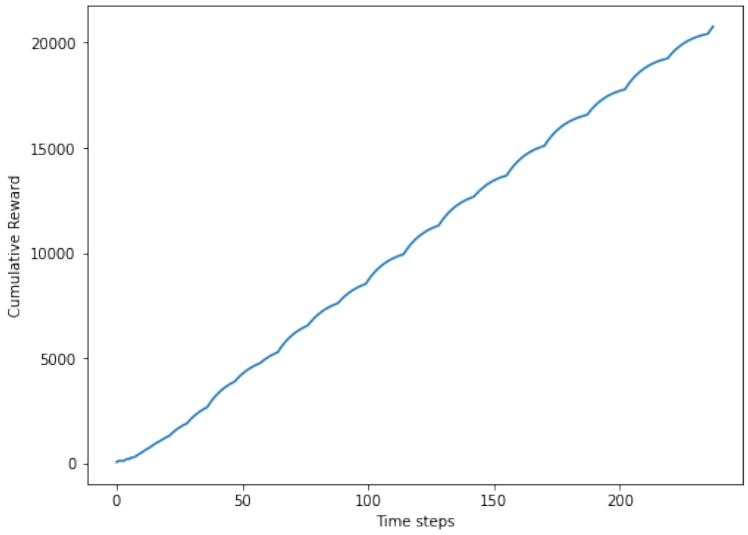

Figure 10: Cumulative reward across time in MVT

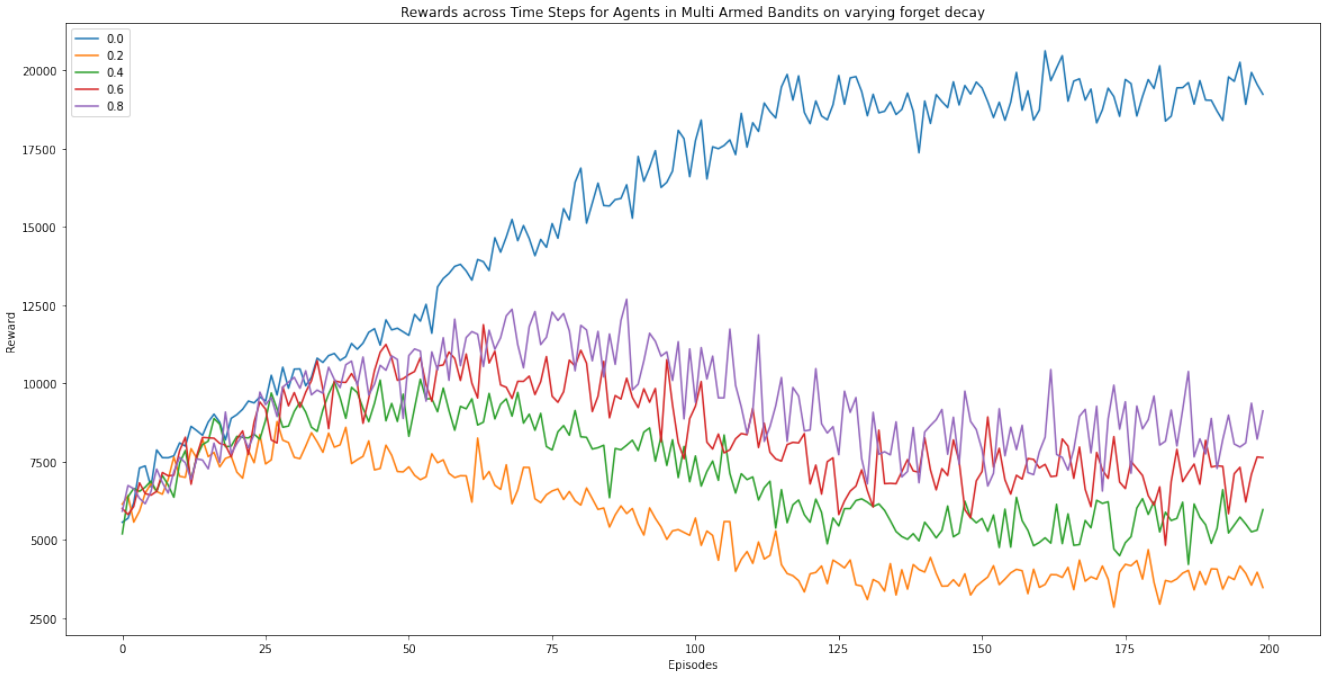

Figure 11: Forgetfulness Model in Decaying Epsilon Greedy Strategy

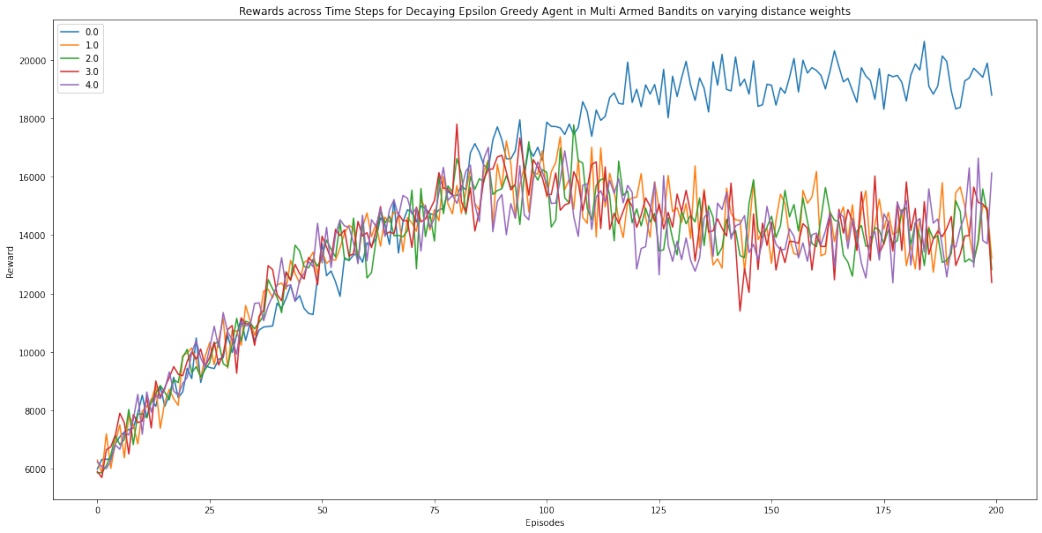

Figure 12: Distance Cost in Decaying Epsilon Greedy Strategy

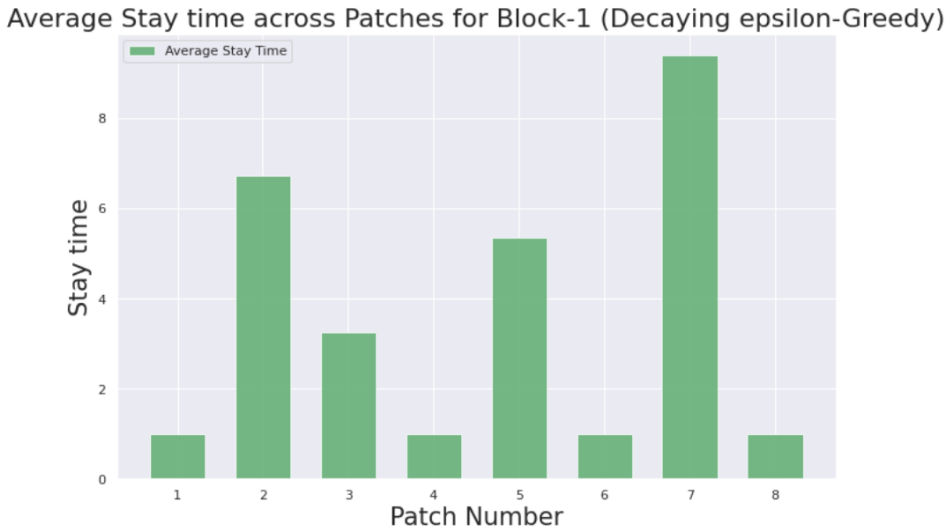

Figure 13: Average Stay Time for Decaying Epsilon Greedy Strategy in Block 1

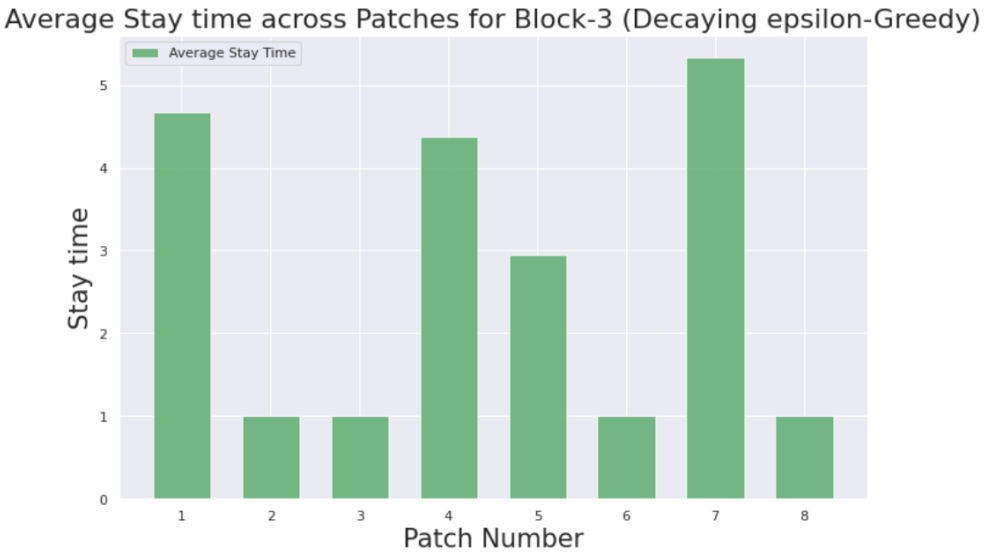

Figure 14: Average Stay Time for Decaying Epsilon Greedy Strategy in Block 3

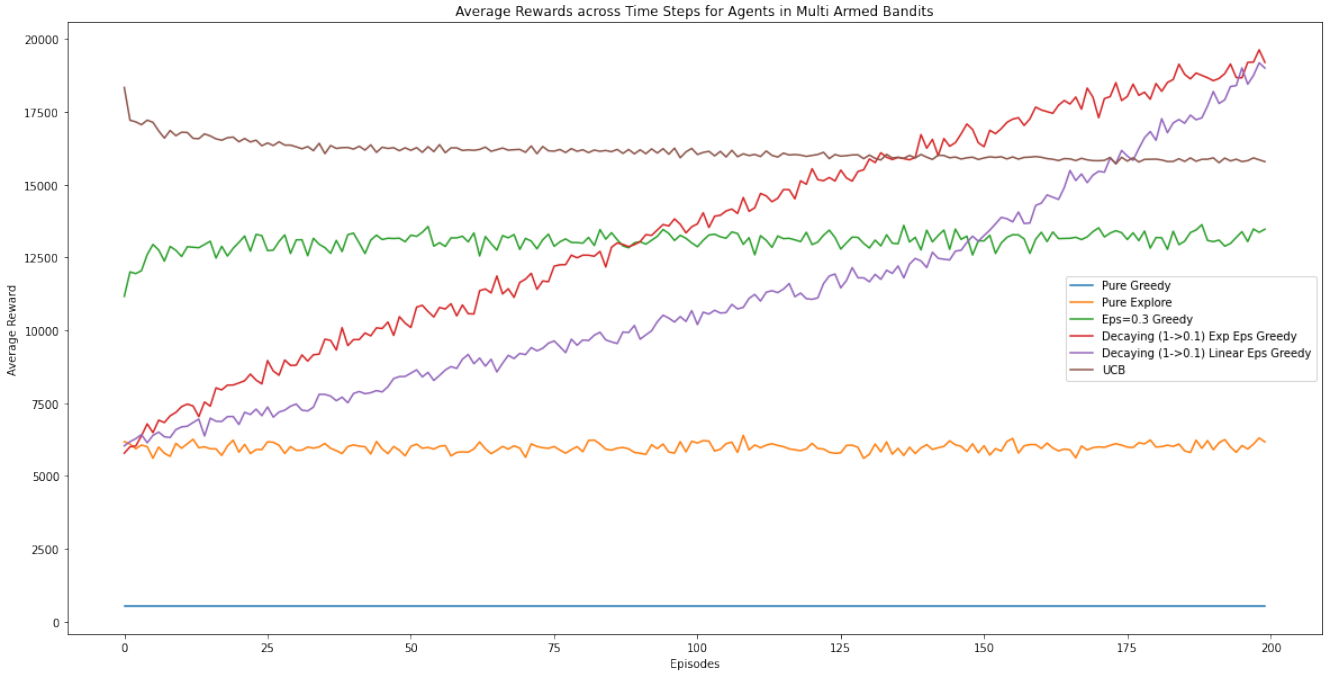

Figure 15: Rewards across Training Episodes for MAB Agents (averaged)

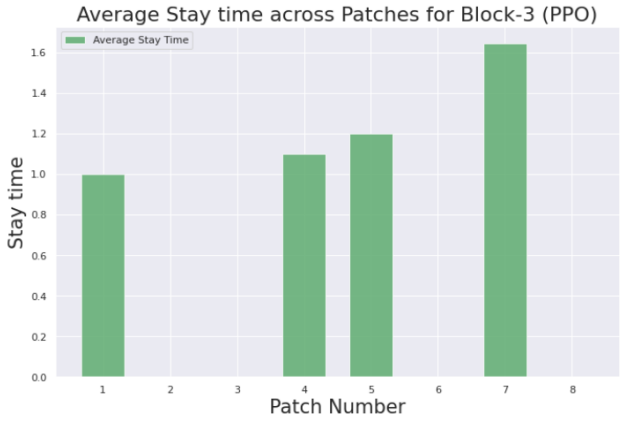

Figure 16: Average Stay Time for PPO in Block 3

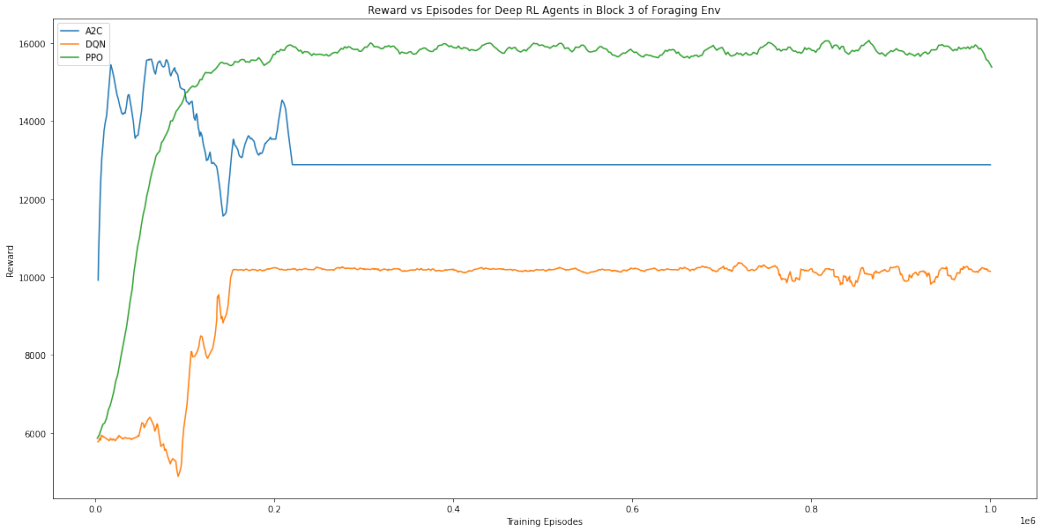

Figure 17: Rewards across training episodes for Deep RL Agents

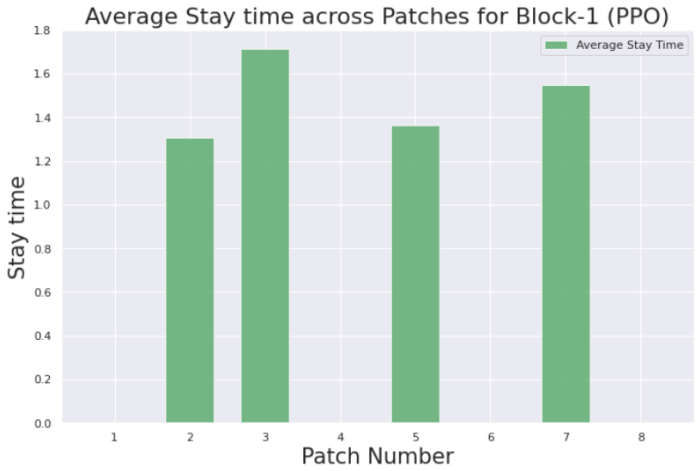

Figure 18: Average Stay Time for PPO Agent in Block 1

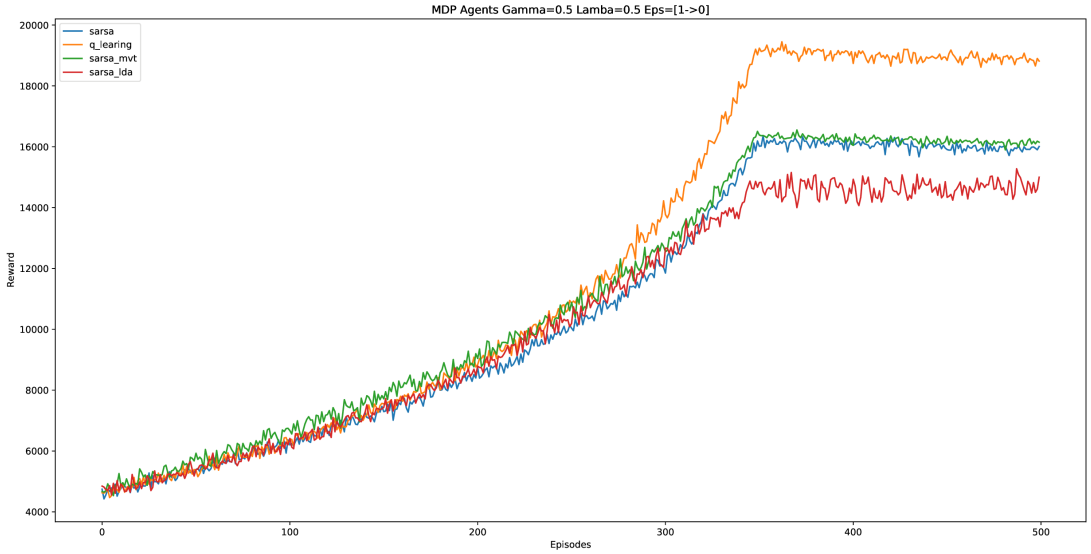

Figure 19: MDP Models Block-1

Figure 20: MDP Models Block-3

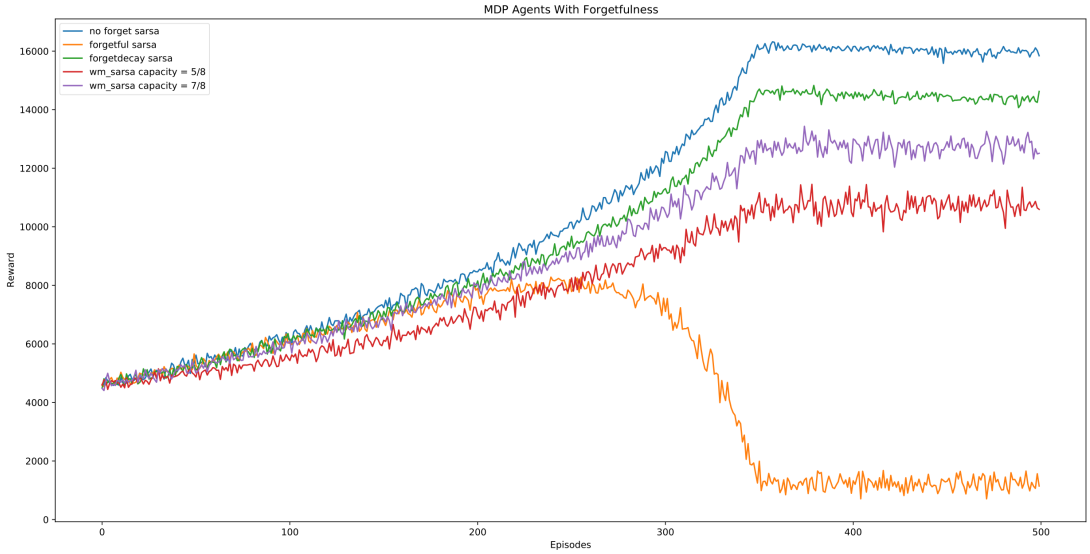

Figure 21: MDP Models with forgetful behaviour Block-1

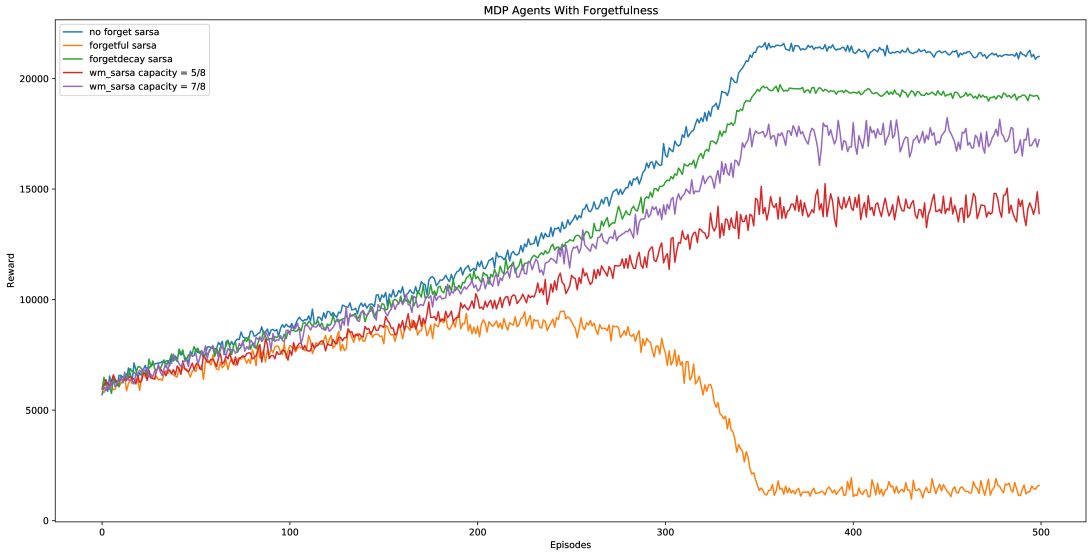

Figure 22: MDP Models with forgetful behaviour Block-3