Assignment 1: Exploration

Hao Chen (s3990788) & Simone de Vos Burchart (s1746995)
Group Number: 88

1 INTRODUCTION

The goal of reinforcement learning is to maximize the overall reward in a decision problem. The 'agent' learns by trying out actions and figuring out which action or series of actions lead to the highest reward. In this paper we will focus on bandits, a one-step decision problem. The agent has only one choice to make, but will do this a number of times with the goal getting the highest expected sum of rewards. Whether the agent repeatedly does the action with the current highest observed reward, or explores new actions to find potentially better rewards, is called the *exploration-exploitation trade-off* [1]. There are multiple ways to balance these two essential sides of learning, and we will evaluate three of them in this paper: ϵ-greedy, optimistic initialization, and upper confidence bounds.

2 ϵ-GREEDY

There are two basic agents: (1) the greedy agent who always chooses the action with the best known reward, and (2) the random agent who always chooses a random action. The greedy agent only exploits, while the random agent only explores. This causes neither agent to learn much. However, combining their strategies and at each time-step choosing between the two agents with a certain probability, will give much better results.

2.1 Methodology

The ϵ-greedy agent uses the greedy strategy with a probability of $1 - \epsilon$, with $\epsilon \in [0, 1]$. With a probability of ϵ it randomly chooses from the other actions.

The action value (mean pay-off) of each action is updated to equal the mean of all received rewards for that action. This update method is called incremental mean update.

2.2 Results

From Figure 1, we can observe that when ϵ = 0.01, the agent learns slower and ends up with lower rewards than an agent with a higher ϵ value. The phenomenon can be attributed to the fact that when ϵ = 0.01, the agent's probability of selecting the action with the highest current reward is 1 - ϵ. So when ϵ = 0.01 the agent has a chance 99% to choose the action with the highest Q-value, and with a mere 1% probability of exploring other actions. This results in the agent being slower to explore and find better actions, with a much flatter curve in Figure 1. However, when ϵ = 0.25, the reward curve does not perform as well as a moderate value of ϵ in the end either. We know the larger ϵ is, the more frequently the agent randomly selects actions. Even if the agent finds the optimal strategy, it will continue to explore suboptimal or even ineffective options, thus affecting the overall rewards. Conversely, when ϵ assumes moderate values (e.g., 0.05, 0.08 and 0.1), the agent requires less time to explore and ultimately achieves higher rewards. It can thus be concluded that the optimal strategy for is found more quickly by the agent if it is allowed to explore to the right extent, without compromising the final performance by over-randomizing the choices.

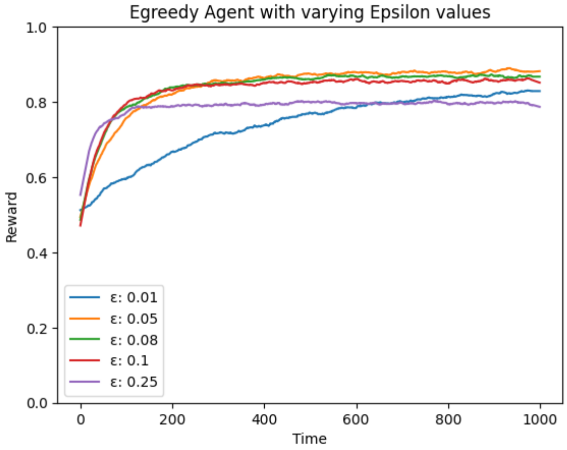

Figure 1: eGreedy Agent with varying ϵ (Epsilon) values

3 OPTIMISTIC INITIALIZATION

Reinforcement learning (RL) agents need to strike a balance between exploration and exploitation when learning from the environment. Especially when there is a cost to the agent's actions, the agent may choose not to explore. In reinforcement learning, we want the agent to choose the optimal action with knowledge of the environment, rather than randomly choosing one action followed by another. One way to encourage exploration without explicitly using randomness is Optimistic Initialization (OI).
We will deliberately set the value estimate of the initial action higher than the realistic expectation. The initially inflated expectation ensures that all actions are fully attempted, helping the agent to learn the exact action values more efficiently. As the agent gains experience and corrects its overly optimistic estimates through incremental updates, it naturally transitions to using actions that have been shown to produce better returns.

3.1 Methodology

We will define an Optimistic Initialization agent that follows a greedy action selection strategy (always select the action with the highest Q-value) but starts with overestimated Q-values for all actions. Over time, as the agent receives rewards, the Q-values are updated and gradually corrected to reflect the actual expected returns.

The Q-value update rule in reinforcement learning with optimistic initialization is given by:

$$Q(a) \leftarrow Q(a) + \alpha \cdot (r(a) - Q(a)) \tag{1}$$

where:

- $Q(a)$ is the estimated value of action a.
- α is the learning rate.
- $r(a)$ is the observed reward for action a.

In our experiments, we keep the learning rate (α) constant at 0.1 while varying the initial Q-values. Our goal is to determine the optimal initial Q-value that leads to the best agent performance in terms of exploration and long-term rewards.

3.2 Results

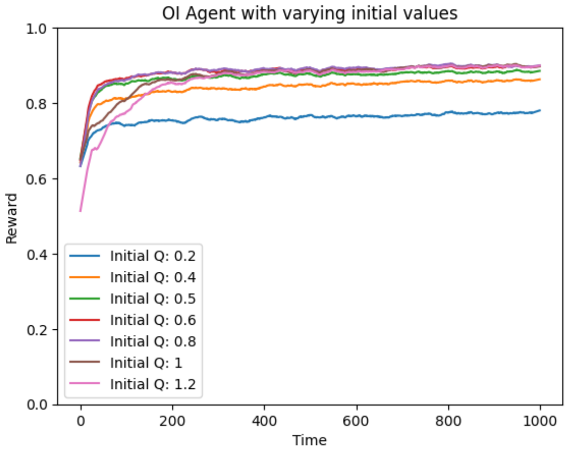

Figure 2: OI Agent with varying initial values, α (learning rate) = 0.1

As observed from Figure 2, when the initial values are set between 0.5 and 0.8, the agent performs best in terms of exploration and long-term rewards. When the initial value is too high (≥ 1), the agent needs more time to correct its Q-values, so it performs poorly in the early stage of exploration. When the initial value is too low (< 0.5), the agent performs poorly in terms of long-term rewards.
According to equation (1), if the initial Q-value is too high, the correction of Q-value will take longer, because the next Q-value is based on the sum of the current Q-value and the difference between the reward from the current action and the current Q-value. The larger this difference, the longer the agent takes to adjust its estimates. This effect is evident when comparing the initial Q-values of 1 and 1.2, where the agent with Q = 1.2 requires more time to explore than the agent with Q = 1.
On the other hand, if the initial Q is too low (e.g. 0.2, 0.4), the agent performs worse in terms of long-term rewards. This is because the Q-value of all actions is initially low, the agent may mistakenly believe that a suboptimal action is optimal after it has tried each action many times. And it will stick to that policy at an early stage of training, causing subsequent learning to suffer.

4 UPPER CONFIDENCE BOUNDS (UCB)

The final policy we implemented is the Upper Confidence Bound, which uses uncertainty to explore in a more targeted manner. Essentially, it selects the action where the reward plus a certain constant times the standard error, is highest. In other words, it selects the action that has the greatest potential with a certain confidence, hence the name Upper Confidence Bounds.

4.1 Methodology

The policy is calculated according to the following equation:

$$\pi_{UCB}(a) = \begin{cases} 1, & \text{if } a = \text{argmax}_{b\in\mathcal{A}}[Q(b) + c \cdot \sqrt{\frac{\ln(t)}{n(b)}}\,] \\ 0, & \text{otherwise} \end{cases}$$

where t is the timestep, $c \in \mathbb{R}^+$ the exploration constant, and $n : \mathcal{A} \longrightarrow \mathbb{N}_0$ the function that maps each action to the number of times it has been tried. As division by 0 is not possible, when $n(a) = 0$ for some action $a \in \mathcal{A}$, we treat the estimate for a as infinity. We tried multiple values for c to evaluate this policy at different learning rates: 0.01, 0.05, 0.1, 0.25, 0.5, and 1.0. Like before, we ran this experiment for 1000 timesteps.

4.2 Results

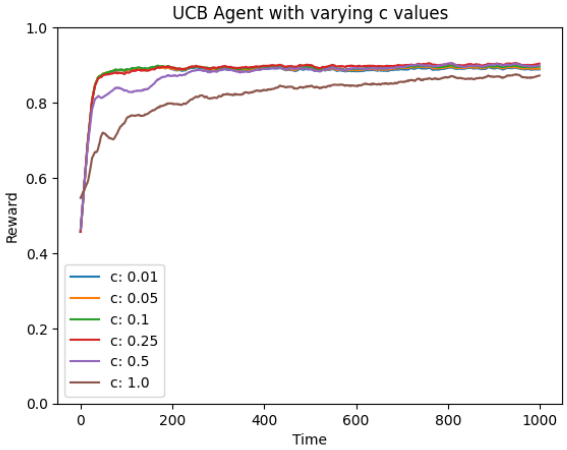

Figure 3: UCB Agent with varying c values

Figure 3 shows a graph of the results of this experiment. With all the values of c that we tried, there was an increase in reward. However, the rate of this increase differed quite a bit, with the reward for $c = 1.0$ being much lower than for all the other timesteps, and $c = 0.5$ having a lower reward than the lower c values, but starting around timestep 300 it is about the same as the lower c values. From around timestep 300, all our values of c less than or equal to 0.5 remained level at around a reward of 0.9. On the other hand, the reward for $c = 1.0$ continues to slowly increase and is only slightly below a reward of 0.9 around timestep 1000, which is not surprising given that a higher c value leads to more exploration and less exploitation.

5 COMPARISON

We have run experiments with different agents, comparing each agent at different parameter values that determine the trade-off between exploration and exploitation. Now we will compare the different agents with each other.

5.1 Methodology

We used the same agents with the same parameters as before:

- ϵ-greedy agent with parameter values $\epsilon \in \{0.01, 0.05, 0.08, 0.1, 0.25\}$
- Optimistic initialization greedy agent with parameter values for the reward initialization in $\{0.2, 0.4, 0.5, 0.6, 0.8, 1\}$ and the learning rate α fixed at 0.1
- Upper confidence bounds agent with parameter values $c \in \{0.01, 0.05, 0.1, 0.25, 0.5, 1.0\}$

We ran the experiment with 1000 timesteps for each agent with each parameter 500 times, and then we took the average per agent-parameter combination.

5.2 Results

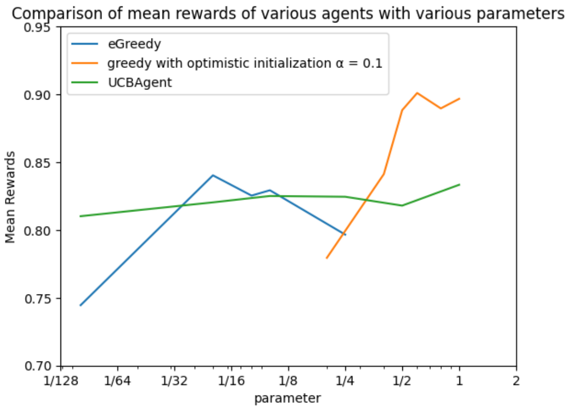

Figure 4: Comparison of various agents

From figure 4, we can observe the following:

- **ϵ-greedy agent**: Performance not really well when ϵ is too small or too large. However, within a moderate range (epsilon = 0.05–0.08), it has been shown to outperform the UCB agent.
- **UCB agent**: Maintains stable performance, with average rewards consistently between 0.8 and 0.85, regardless of the chosen c value.
- **Optimistic initialization agent**: Improves as the initial value increases. However, when the initial value exceeds 0.6, a slight decrease in the average reward is observed, suggesting that excessive initialization may negatively impact performance.

6 DISCUSSION

In this study, we evaluated three reinforcement learning strategies: eGreedy, optimistic initialization, and upper confidence bounds. The treatment of the exploration-exploitation trade-off in exploration problems is investigated.

For the eGreedy agent, our experiments demonstrated that selecting an appropriate ϵ value is crucial for balancing exploration and exploitation. A very small ϵ led to a longer lasting suboptimal strategy due to insufficient exploration, while a very large ϵ leads to the agent choosing actions randomly. The optimal ϵ range (0.05-0.08) allowed the agent to explore sufficiently while still favoring high-reward actions, resulting in better overall performance.

The optimistic initialization approach showed that setting an initial Q value with an appropriate range (0.5-0.8) improved early-stage exploration while maintaining strong long-term performance. If the initial Q-value was set too high (e.g. 1.2), the agent required more time to correct its estimates, so that agent needs a long time to get the optimal action. Conversely, too low an initial Q-value (<0.5) caused premature convergence to suboptimal actions, negatively affecting long-term rewards.

The UCB method showed stable performance over different values of the exploration constant. Lower values of c led to faster convergence with moderate exploration, while higher values (e.g. 1.0) prioritized exploration at the cost of initial performance. Interestingly, despite the initial slow learning, the agent continued to improve over time and eventually reached performance levels close to those of agents with other values.

The present study has only been tested on the three strategies separately. It would be advisable for future research to combine some of these strategies in order to ascertain whether this results in superior outcomes. Moreover, in this study, the learning rate for the OI agent was set at 0.1. Future work could explore the impact of varying learning rates to determine how different values affect the agent's convergence speed and overall performance.

7 CONCLUSION

In summary, our study shows that each of these strategies has its own advantages and disadvantages. When the data is uncertain or a fast suboptimal solution is preferred, the eGreedy strategy can help strike a balance between exploration and exploitation. If the goal is to find the optimal solution and the per-step cost is not excessively high, the optimistic initialization (OI) agent may be a better choice, as it enables the agent to identify the optimal action through multiple trials. The UCB method provides a structured approach to exploration, making it particularly advantageous in scenarios where minimizing unnecessary exploration costs is crucial. Our future recommendation is to try changing ϵ over time, e.g. starting with $\epsilon = 0.25$ (which started the best in Figure 1) and gradually lowering it when there has been sufficient exploration. This will increase the mean rewards while not lowering the reward it converges to.

8 APPENDIX

8.1 Different learning rate in OI Agent

In our previous experiments, the optimistic initialization (OI) agent used a fixed learning rate (α=0.1). To investigate the impact of different learning rates, we tested the agent with varying values of $\alpha = \{0.01, 0.05, 0.1, 0.2, 0.3\}$.

Since initial Q-values play a crucial role in learning behavior, we selected two different initial values, 0.6 and 0.8, and plotted the reward progression over time for both cases. The purpose of this comparison was to analyze how different learning rates interact with different initial Q-values, ultimately affecting the agent's learning efficiency.

As shown in the plots in Figure 5 and 6, when the initial value is 0.8 (Figure 6), the agent performed best when α was in the range of 0.05 to 0.2. Lower α (e.g. 0.01) resulted in slower convergence, while higher α (e.g. 0.3) introduced more fluctuation in rewards.

As shown in Figure 6 and 5, when α=0.01, the agent with an initial value of 0.6 performed worse than the agent with an initial value of 0.8. The agent learned more gradually, but high α values (0.2 and 0.3) helped achieve faster adaptation compared to lower α.

These results indicate that the optimal learning rate depends on the initial Q-value. If the initial Q-value is higher, a moderate learning rate (α between 0.05 and 0.2) is preferable to maintains stability while enabling efficient updates. However, when the initial Q-value is lower, slightly higher α values may help the agent adapt faster.

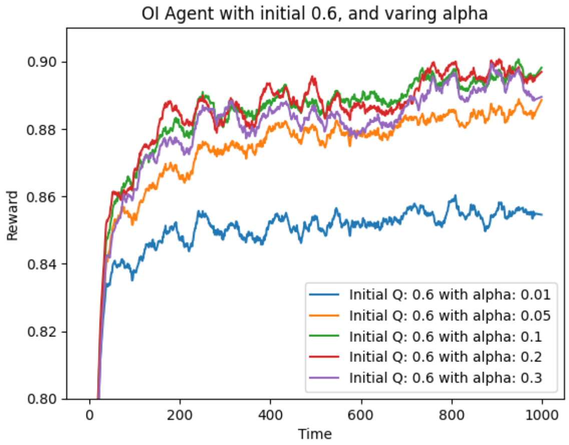

Figure 5: OI Agent with initial 0.6, with varying α

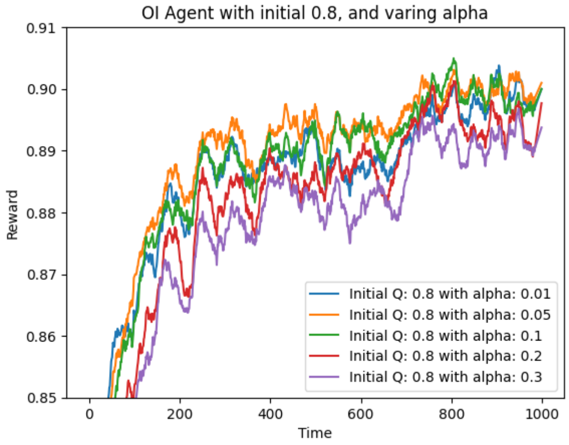

Figure 6: OI Agent with initial 0.8, with varying α

8.2 OI agent with ϵ-Greedy policy

Given the advantages of both optimistic initialization (OI) and eGreedy strategies, we explored whether combining these two approaches could improve performance in reinforcement learning. Instead of following a purely greedy selection strategy in the OI agent, we introduced an eGreedy mechanism, allowing the agent to explore with probability ϵ while still leveraging optimistic initialization.

To evaluate this approach, we tested different initial Q-values and constant ϵ value, keeping the learning rate α=0.2 and ϵ value= 0.05 (which is best performance of eGreedy Agent) constant. Our goal was to determine whether this hybrid strategy could enhance learning efficiency and long-term reward accumulation.

The results, however, were contrary to our expectations—the combination of OI and eGreedy did not lead to improved performance. The experimental results indicate that the OI + eGreedy agent consistently underperformed compared to the standard OI or eGreedy agent alone. As shown in Figure 7, when the initial Q-values were 0.5 and 0.6 with the eGreedy strategy performed worse than the agent that always selected the action with the highest current Q-value (i.e., the standard OI agent).

This performance degradation is primarily due to two factors:

- Inefficient correction of optimistic initial values: Since the eGreedy agent continuously explores, it fails to refine its Q-values effectively, leading to prolonged instability in the learning process..
- Excessive exploration hindering convergence: While eGreedy encourages exploration, in this case, it prevented the agent from fully exploiting optimal actions, causing it to settle on suboptimal policies.

These findings suggest that while both OI and eGreedy perform well individually, their combination disrupts learning dynamics rather than enhancing them.

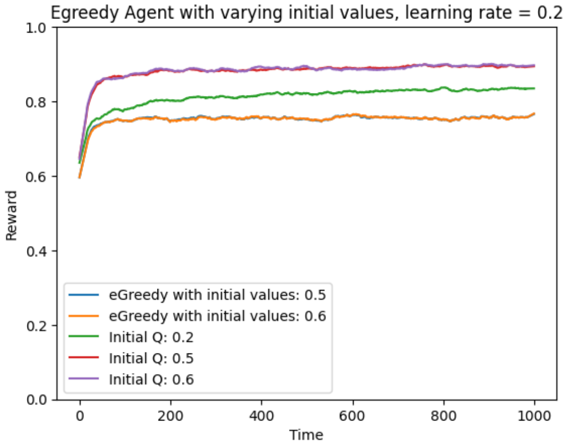

Figure 7: eGreedy Agent with varying initial value, α = 0.2

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction.* MIT press, 2018.