Assignment 1

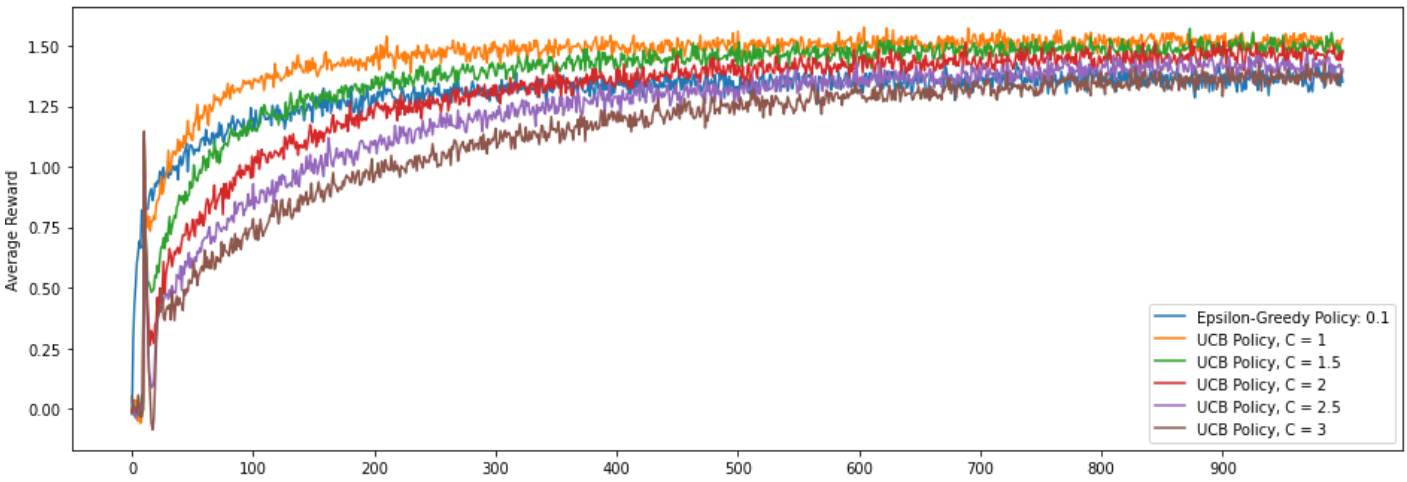
February 6, 2022

1. Write $250 - 500$ words on the areas you wish to explore with reinforcement learning. Be sure to include why particular areas/subfields interest you.

   In these few semesters, I have learned many algorithms in the field of computer vision, such as regression, neural networks and so on. However, I don't think these algorithms can be called intelligence, I think they're just doing classification. During these weeks learning, we created agents just like humans, they will do the exploration part and remember actions and states with higher rewards. I believe this is closer to the basic principles of human intelligence. Just like we were very curious about the world when we were young. We grow up because we can remember which behaviors are dangerous and which ones can be rewarded.

   For reinforcement learning, the direction I am most interested in is whether we can use reinforcement learning techniques to reduce over fitting. When I was learning image classification, we gave each training object a true label. Consider if we don't give the agent a true label, but for each batch, give it a total score, such as 90/100. Just like the TOEFL test. I believe this can reduce over fitting. However, it is hard for me to define a loss function for this idea. Without the loss function, I also cannot update the gradient in the back propagation. Therefore i was wondering if i can train two agent, one is teacher and the other is student. Then we can simulate the process of teaching. We pre-train the teacher for some epoch and let the teacher check the student's weight. Then the student will do exploration part and re-update teacher's weight. I believe this can reduce the student's training time and over fitting. This direction interest so much is because this process simulate how human pass knowledge and generate new knowledge. I think this semester's learning can help me better understand this idea.

2. Recreate Figure 2.3 in the text. Run some experiments of your own to try and get better performance, and plot those alongside these results in a second graph. Use assignment1.pynb on OWL (Week 1) as a base and edit to complete the assignment.



   I have chosen $c = 1, c = 1.5, c = 2, c = 2.5 and c = 3$, among all of these experiments, $c = 1$ gets the best performance on average reward. Many of these has a experiments has a prominent spike on the 11th play. I think this is because we perform UCB action on the 10-armed testbed. After all 10 arm explored, the UCB police will start to choose the optimal policy hence the average reward at 11th play will increase a lot.

3. Explain in simple language how UCB enables exploration. You can do this in the comment sections in the relevant code.

   (a) $\epsilon -$ greedy divide the selection process into two phases: Exploration and Exploitation. It will explore with the same probability ($\frac{\epsilon}{N}$) while exploring.

(b) UCB define a bias $= c[\sqrt{\frac{\log(t)}{N_t(a)}}]$. If the number of an action has been chosen is small, then the bias term got bigger. As the time goes on, it will more likely to encourage the agent to explore on other actions.