Bonus Question 1

Epsilon-Greedy

Advantages: easy to implement; can be easy to apply in complex settings
Disadvantages: sensitive to alpha, epsilon and initial Q estimates

UCB

Advantages: less sensitive to hyper-parameters compared to epsilon-greedy, learning rate can be pretty big and tuning the parameter c within some range won't hurt the performance too much.
Disadvantages: difficult to apply in more complex settings; uses much space because we have to store table of the number of selected actions.
*The plot of the UCB agent's performance is included in the zip file.

Bonus Question 2

$$Q_n = \prod_{i=1}^{n}(1-\alpha_i)Q_0 + \sum_{i=1}^{n}\alpha_i\prod_{j=i+1}^{n}(1-\alpha_j)Ri$$

Bonus Question 3

The spike will appear in almost every experiment, which is decided by the property of optimistic initialization. In the early steps because we can systematically explore, so the agent has a good chance to find the best action; however, because the actual reward of the optimal action is lower than the initial estimation, with increasingly choosing the optimal action, the Q value drops and the agent will choose other actions of higher Q value. Although the Q value of the optimal action is more accurate than the estimation of other acitons in this case, because we are using optimistic initialization, we are forcing the agent to explore more and therefore it will has a spike at the early stage.