ECE 289A - An Introduction to Reinforcement Learning HW#1

Ahmed H. Mahmoud

October, 10th 2017

Figure 1 shows the regenerated plot from Figure 2.4 in [1]. The figure shows the comparison between *upper confidence bound* (UCB) and ϵ-*greedy* method (with ϵ =0.1) on K-armed bandit problem with $K = 10$. The plot was generated by averaging the result of 2000 randomly generated K-armed bandits over 1000 steps. We notice that there is a spike on the 11th step on the UCB

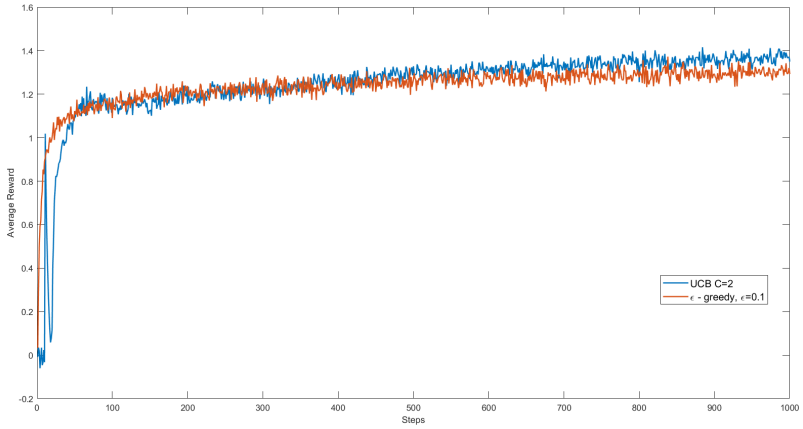

Figure 1

curve. Since the selection of the next action is based on the following equation

$$A_t = argmax(Q_t(a) + c\sqrt{\frac{log(t)}{N_t(a)}})$$

where A_t is the action at time t, $Q_t(a)$ is the estimate of action a at time t, c is the UCB parameter that controls the degree of exploration, and $N_t(a)$ is the number of times action a has been taken up to time t. At the beginning, all action have $N_t(a) = 0, \forall a \in K$. Thus, the second term ($c\sqrt{\frac{log(t)}{N_t(a)}}$) is maximum and the action that has not been chosen yet will have a maximum value and will be selected (with random tie breaker). This is will last for first K steps ($K = 10$) which guarantees that all actions will be explored first. On the 11th step, all actions will have same value for the second term and only the best action will have a maximum estimate and thus it will be chosen which explains the spike at the 11th step.

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, Second edition, 2017.