Lecture by Prof. Eleni Vasilaki

Exercise 1

Show that the empirical mean $Q_k = \frac{r_1+...+r_k}{k}$ can be iteratively updated by $\Delta Q = \eta(r - Q)$. What is the optimal η for the most accurate estimation of Q_k?

Let us express Q at time step $k+1$ as a function of Q at time step k:

$Q_{k+1} = \frac{r_1+...+r_{k+1}}{k+1} = \frac{k(r_1+...+r_k)}{k(k+1)} + \frac{r_{k+1}}{k+1} = \frac{k}{k+1}Q_k + \frac{r_{k+1}}{k+1} = Q_k - Q_k + \frac{k}{k+1}Q_k + \frac{r_{k+1}}{k+1}$

$Q_{k+1} = Q_k - \frac{1}{k+1}Q_k + \frac{1}{k+1}r_{k+1} = Q_k + \frac{1}{k+1}(r_{k+1} - Q_k)$

Thus $\eta = \frac{1}{k+1}$.

Exercise 2

In a 2-armed bandit problem, the two possible actions have expected values $Q^*(a_1) = 1$ and $Q^*(a_2) = 2$. Upon acting a_1, the agent always gets reward $r = 1$. Upon acting a_2, it receives an integer reward uniformly distributed in the interval [-3,7].

1. *Confirm that $Q^*(a_2) = 2$.*
2. *If you repeat this experiment many times, calculate the expected reward of the agent in the long run if the agent chooses an action in the following way: Perform once the action a_1, once the action a_2 and then always chose the action that first yielded the better outcome (if equal choose a_1).*
3. *Calculate the expected reward in the long run if ϵ-greedy action selection is used. For which ϵ would the expected result be better than the previous strategy?*

1. $Q^*(a_2)$ is the expected reward, i.e. $\sum_{r=-3}^{r=7} rp(r)$. Since reward is an integer uniformly distributed in [-3,7], it can be one of the values -3,-2,-1,0,1,2,3,4,5,6,7 i.e. 11 different option with probability for each 1/11. Thus $Q^*(a_2) = \frac{\sum_{r=-3}^{r=7} r}{11} = 2$. Note: You could also calculate this by finding the middle point of the space [-3,7], i.e. $Q^*(a_2) = \frac{7-(-3)}{2} + (-3) = 2$.
2. In this scenario, if the first sample of a_2 is smaller or equal to 1 (which has a probability of p=5/11), the agent will always choose a_1, which is the worst of the two actions. Thus, with probability 1-p=6/11 the optimal action, a_2 will be chosen.

   $E(r) = pQ^*(a_1) + (1-p)Q^*(a_2) = \frac{5}{11}1 + \frac{6}{11}2 = 17/11 < 2.$
3. If ϵ-greedy action selection is used, the agent will discover the true values of the two actions. It will therefore chose the best action a_2 with probability $p(a_2) = 1-\epsilon$ and action a_1 or a_2 with probability $p = 0.5\epsilon$. Thus:

   $E(r) = (1-\epsilon)Q^*(a_2) + \frac{\epsilon}{2}Q^*(a_2) + \frac{\epsilon}{2}Q^*(a_1) = 2 - 0.5\epsilon$

   We wish to find values of η for which should be higher than the reward calculated in the previous scenario i.e. $2-0.5\epsilon > 17/11$. We conclude that values of $\epsilon < 10/11$ will result in a higher expected reward, and thus ϵ-greedy is a better scheme in this case.