ECE 493, Spring 2020, Assignment 1
Due: Friday June 19, 11:59pm

Submit to the UWaterloo Crowdmark site using the link you received via email. Your answer can be handwritten converted to an electronic file by and scanner or camera; or the answers can be typed up in a word processor or LaTeX and submitted as a pdf.

1 Multi-Armed Bandits

1. Consider a k-armed bandit problem with k = 4 actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of Q1(a) = 0, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1$, $R_1 = 1$, $A_2 = 2$, $R_2 = 1$, $A_3 = 2$, $R_3 = 2$, $A_4 = 2$, $R_4 = 2$, $A_5 = 3$, $R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

Table 1: k-Armed Bandit Actions

Step	Action	Reward	Action with Largest Reward	ϵ action	Greedy action
1	1	1		Maybe	Maybe
2	2	1	1	Yes	No
3	2	2	1,2	Maybe	Maybe
4	2	2	2	Maybe	Maybe
5	3	0	2	Yes	No

Note: All actions can potentially be an ϵ action, as an ϵ action choice is independent of action values (meaning it can pick the best action), " ...instead select randomly from among all the actions with equal probability, independently of the action-value estimates ".

2. Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 10 and 20 with probability 0.5 (case A), and 90 and 80 with probability 0.5 (case B). If you are not able to tell which case you face at any step, what is the best expected reward you can achieve and how should you behave to achieve it? Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expected reward you can achieve in this task, and how should you behave to achieve it?

For the first part of the question, it does not matter which initial bandit is chosen. After choosing an initial bandit, the greedy approach should be employed (exploiting and never exploring). If the greedy approach is taken, 50% of the choices will lead to un-optimal cases, and 50% of the choices will lead to optimal cases, but the reward difference between optimal and sub-optimal is the same for action 1 and action 2.

If in each case the reward difference between the two bandits where unequal, the best approach would be to average the action values, and choose the greedy action from the average of the actions value.

For example, if we iterate on $2x$ number of actions, we get the following rewards depending on the initial bandit chosen.

- A1 initial bandit: x*10 + x*90 = x*100
- A2 initial bandit: x*20 + x*80 = x*100

For the second part of the question, if we know whether the environment is in state A or state B, the first set of actions should be spent exploring the environment to completely map out a single case, either A or B. This will take at most 3 actions. Once a single case has been mapped, the agent can choose the greedy action for that state, and the opposite action for the second case.

For example, if we find that action 1 is optimal in case A, for all case A environments the agent will choose action 1, and action 2 for case B.