School of Engineering and Applied Science
The George Washington University

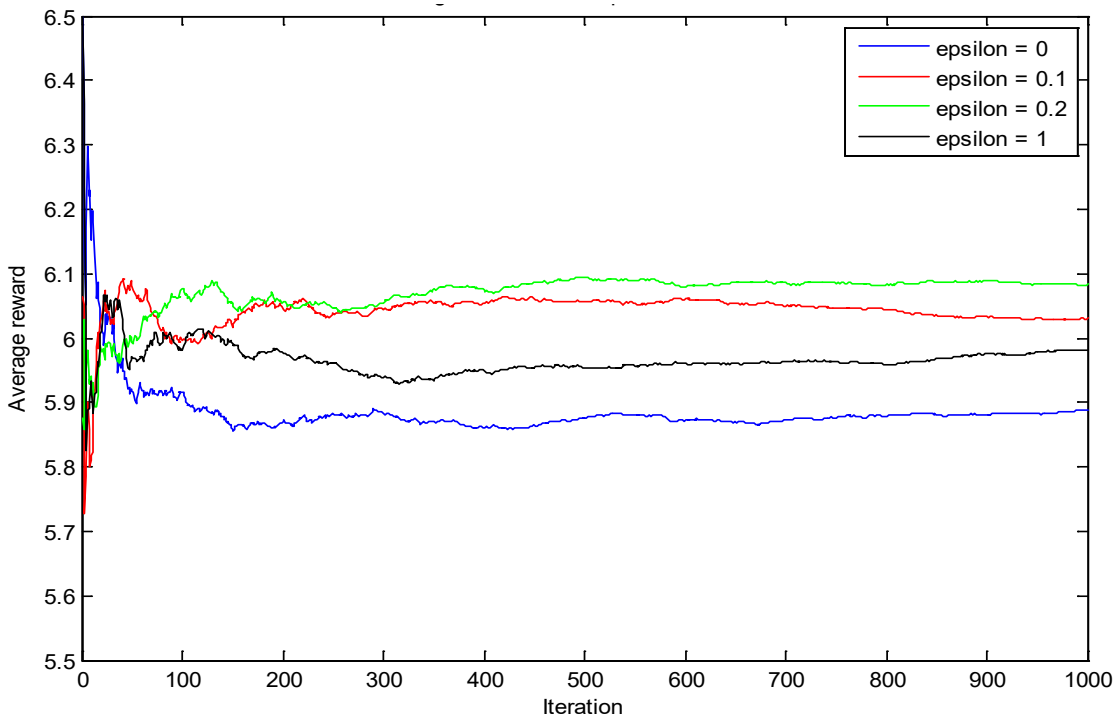
ECE 6045 **Homework #1** **Spring 2020**

Think of an agent that plays a 2-armed bandit, trying to maximize its reward. In each step, the agent selects one of the levers, and is given some reward according to the reward distribution of that lever.

Assume that reward distribution for the first lever is a Gaussian with $\mu_1 = 5, \sigma_1^2 = 10$, and for the second lever is a binomial Gaussian with $\mu_{21} = 10, \sigma_{21}^2 = 15, \mu_{22} = 4, \sigma_{22}^2 = 10$, which means that the resulting output will be uniformly probable from these two Gaussian distributions (See http://en.wikipedia.org/wiki/Mixture_distribution). Implement the ϵ-greedy action selection policy and compare the average reward in the first 1000 steps in a graph. In order to have smoother graphs, each curve should be the average of 100 different runs.

a) Report the action-values for various learning rate α and ϵ-greedy parameter. You need to use $Q(k+1) = Q(k) + \alpha(\, reward - Q(k)\,)$ for updating the action values. For the learning rates, consider the following values: $\alpha = 1, \alpha = 0.9^k, \alpha = \frac{1}{\alpha^k}$ and $\alpha = \frac{1}{k}$, and for the policy, use the ϵ-greedy method with $\epsilon = 0,\ 0.1,\ 0.2, 0.5$.
b) For a fixed $\alpha = 0.1$ and $\epsilon = 0.1$, use the following optimistic initial values and compare the results: $Q_0 = [0\ 0]$, $Q_0 = [5\ 5]$, $Q_0 = [10\ 10]$.

Note: For each question, you should provide the results, your explanations of the acquired results and your source codes. For each choice of α, you need to provide plot and a table similar to the following ones:



Epsilon	Final average of 100 runs	Final standard deviation of 100 runs	Average of Final $\hat{Q}_1$	Average of Final $\hat{Q}_2$
ε = 0 (greedy)				
ε = 0.1				
ε = 0.2				
ε = 0.5 (random)				

Q_0	Final average of 100 runs	Final standard deviation of 100 runs	Average of Final $\hat{Q}_1$	Average of Final $\hat{Q}_2$
Q_0 =[0 0]				
Q_0 =[5 5]				
Q_0 =[10 10]				
Q_0 =[7 7]				