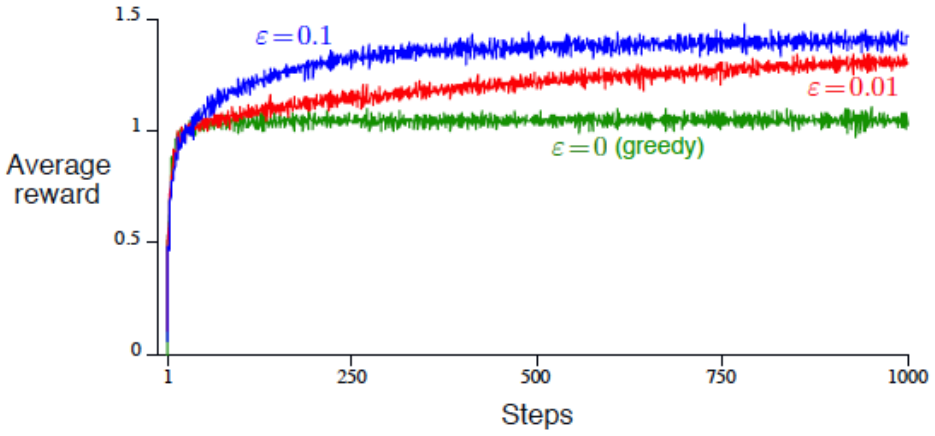
# Homework 1

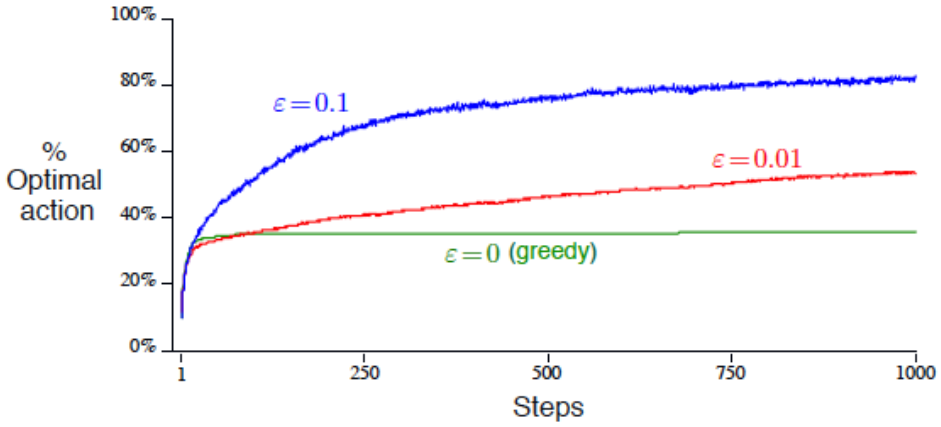
Due date: 2nd Apr., 2019

1. **(Matlab Exercise for 10-armed Testbed)** Consider a set of 2000 randomly generated 10-armed bandit problems. For each bandit problem, the action values $q_*(a)$ for $a$=1,…,10 were selected according to a normal (Gaussian) distribution with mean 0 and variance 1. Then, when a learning method applied to that problem selected action $A_t$ at time step $t$, the actual reward $R_t$ was selected from a normal distribution with mean $q_*(A_t)$ and variance 1. For any learning method, we can measure its performance and behavior as it improves with experience over 1000 time steps (one run). Repeating this for 2000 independent runs, each with a different bandit problem, we then obtain the average behavior.

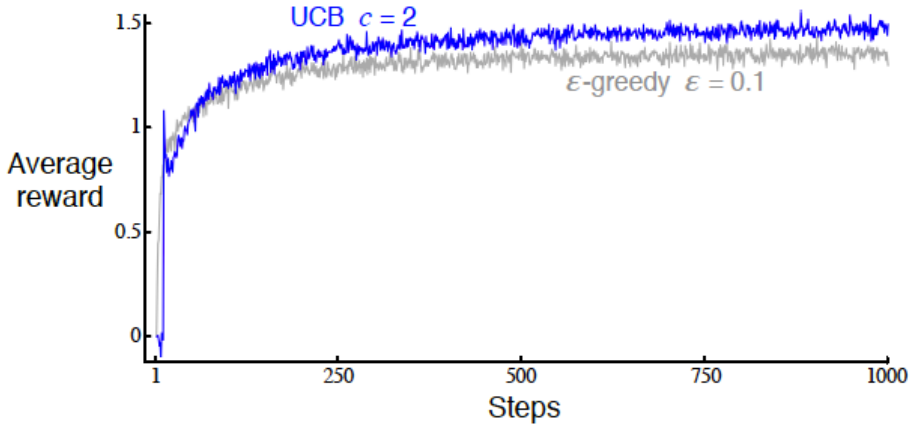
   A. (35 points) Plot the following figure with curves for the $\varepsilon$-greedy method ($\varepsilon$=0.1 and 0.01) and the greedy method ($\varepsilon$=0).




   B. (35 points) Plot the following figure with curves for the $\varepsilon$-greedy method ($\varepsilon$=0.1 and 0.01) and the greedy method ($\varepsilon$=0).



C. (30 points) Apply the upper-confidence-bound (UCB) action selection and plot the following figure with curves for the $\varepsilon$-greedy method ($\varepsilon$=0.1) and the UCB method ($c$=2).




**Please submit (single or multiple) .m file to ecourse. The figures should be able to automatically generated after compiling. Please provide sufficient elaboration for each code/variable/parameter statement.**