ME 537
Learning Based Control
Fall 2017
HW #3: Reinforcement Learning
Due 10/30

1- Use your favorite programming language to implement a simple reinforcement learning algorithm to solve the N-armed bandit problem. Note, this system does not have "states" so you will be using an "action-value" learner, where each action directly results in a reward.

Consider the case with five actions, where the reward for each action is drawn from a Gaussian distribution with the following mean and variances (for example, when you take action A1, you get a reward with mean=1 and variance=5):

Action	Mean	Variance
A1:	1	5
A2:	1.5	1
A3:	2	1
A4:	2	2
A5:	1.75	10

Compare the performance of selecting actions using a greedy and e-greedy algorithm, when an episode is 10 time steps. Each time step consists of the learner picking an action and getting the corresponding reward. Repeat the experiment with episodes of 100 time steps. Did the results change? Provide justifications to your answer.

2- Consider a 5x10 gridworld. There is a door at the lower right hand side of this grid (red). The agent starts at a random location and has five actions (move in four directions or stay in place). There is a reward of 100 (red box) to exit through the door and a reward of "-1" for being in every other cell. Use the **EXACT SAME** RL algorithm devised in part 1. Use the e-greedy action selection for episodes of 20 steps. How did the algorithm perform? What are the problems?

3- Implement a Q-learning algorithm and use it to solve the same, 20 step gridworld problem. How did the algorithm perform? How fast did it learn? How did solutions compare to the simple action value algorithm tried for problem 2? What was the key difference between the two algorithms? Discuss the implications of your results.