ME 537
Learning Based Control
Fall 2010
HW #3: Reinforcement Learning
Due 11/5/2010 and noon

1- Use your favorite programming language to implement a simple reinforcement learning algorithm (action value function) to solve the N-bandit problem.

Consider the case with five actions ,where the mean and variance of each action is given by:

A1: 1,5
A2: 1,1
A3: 2,1
A4: 2,1
A5: 0,10

Compare greedy, e-greedy and softmax action selection, for when an episode is 10 and 100 time steps.

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 nothing otherwise. Use the **exact same** RL algorithm devised in part 1 on this problem. How does it work? What are the problems?

Now, implement a Q-learning algorithm on the same problem. How well does it work? How fast does it learn? How do solutions compare to the simple action value algorithm?

Discuss the implications of your results.