**CS6700: Reinforcement Learning** **Spring 2017**

# Written Assignment 2

*Lecturer: Balaraman Ravindran* *TA: Aravind S, JP Sagar, Sahil S, Subhojyoti M*

1. Consider the following grid world task. The environment is a $10 \times 10$ grid. The aim is to learn a policy to go from the start state to the goal state in the fewest possible steps. The 4 deterministic actions available are to move one step up, down, left or right. Standard grid world dynamics apply. The agent receives a reward of 0 at each time step and 1 when it reaches the goal. There is a discount factor $0 < \gamma < 1$. Formulate this problem as a family of bandit tasks. These tasks are obviously related to one another. Describe the structure of the set up and the rewards associated with each action for each of the tasks, to make it perform similarly to a Q-learning agent.

2. Consider a bandit problem in which you know the set of expected payoffs for pulling various arms, but you do not know which arm maps to which expected payoff. For example, consider a 5 arm bandit problem and you know that the arms 1 through 5 have payoffs 3.1, 2.3, 4.6, 1.2, 0.9, but not necessarily in that order. Can you design a regret minimizing algorithm that will achieve better bounds than UCB? What makes you believe that it is possible? What parts of the analysis of UCB will you modify to achieve better bounds? Note that I am not asking you for a complete algorithm or analysis, only the intuition.

3. Define a bandit set up as follows. At each time instant for each armof the bandit we sample a reward from some unknown distribution. Now the agent picks an arm. The environment then reveals all the rewards that were chosen. Regret is now defined as the difference between the best arm at that instant and the one chosen summed over all times steps. Would the existing algorithms for bandit problems work well in this setting? Can we do better by taking advantage of the fact that all rewards are revealed? For e.g., exploration is not an issue now, since all arms are revealed at each time step.

4. The results shown in Figure 2.3 (of course text book uploaded in moodle) should be quite reliable because they are averages over 2000 individual, randomly chosen 10-armed bandit tasks. Why, then, are there oscillations and spikes in the early part of the curve for the optimistic method? In other words, what might make this method perform particularly better or worse, on average, on particular early steps?

5. If the step-size parameters, $\alpha_n$, are not constant, then the estimate $Q_n$ is a weighted average of previously received rewards with a weighting different from that given by (eq 2.6 of course text book). What is the weighting on each prior reward for the general case, analogous to (eq 2.6 of course text book), in terms of the sequence of step-size parameters?