E1-277 Reinforcement Learning: Mid-Term Exam 1

Total Marks: 15; Answer all the questions

1. Consider a multi-armed bandit problem with a total of K arms. Assume that arm i when pulled gives a reward with distribution $N(\mu, \sigma_i^2)$, $i = 1, \ldots, K$. Notice that the mean of the rewards (μ) is the same for all arms but they differ in the reward variance (σ_i^2). Define $Q_t(a), t \geq 1$ as before, i.e.,

$$Q_t(a) = \frac{\text{sum of rewards when } a \text{ is taken prior to } t}{\text{number of times } a \text{ is taken prior to } t}.$$

Our goal in this problem is to pick the arm that has the least reward variance. Write down a procedure for this after devising a suitable "ϵ-greedy" strategy as well as an incremental update algorithm for the reward variance. (3)

2. Recall the Bellman equation for optimality for the stochastic shortest path problem:

$$J(i) = \min_{u \in A(i)} \sum_{j=1}^{n} p_{ij}(u)(g(i,u,j) + J(j)), \ i = 1, \ldots, n,$$

for which the solution we know is $J^*(i)$. Define now

$$Q^*(i,u) = \sum_{j=1}^{n} p_{ij}(u)(g(i,u,j) + J^*(j)), \text{ with } u \in A(i), \ i = 1, \ldots, n;$$

$$\tilde{Q}(i,u) = \sum_{j=1}^{n} p_{ij}(u)(g(i,u,j) + \tilde{J}(j)), \text{ with } u \in A(i), \ i = 1, \ldots, n,$$

for some functions $\tilde{Q}(i,u)$ and $\tilde{J}(i) = \min_{u \in A(i)} \tilde{Q}(i,u)$.

   (a) Write down a Bellman equation (i.e., a fixed point equation) in terms of state-action values $Q(i,u)$, $i \in S$, $u \in A(i)$ instead of values $J(i), i \in S$? (1.5)

   (b) Show whether or not $Q^* = \tilde{Q}$? (1.5)

3. Consider a house with four rooms that are identified with the states of a Markov decision process and numbered as states 1,2, DG and UG, respectively, see Figure 1. The states DG and UG are goal states. If the process enters into either of these states, it just stays there and does not come out of them. Here DG is the desirable goal state and UG is the undesirable goal state. The transition dynamics is as follows: There are two actions A_1 and A_2 that are feasible in state 1 and only an action A feasible in state 2. When action A_1 is chosen in state 1, the process moves to state 2 with probability 0.7 giving a reward of 0 and moves to state DG with probability 0.3 and gives a reward of 1. When action A_2 is chosen in state 1, the process remains in state 1 with probability 0.5 giving a reward of 0 and to DG with probability 0.3 giving a reward of 1 and to state UG with probability 0.2 giving a reward of 0. When in state 2, upon selecting action A, the process moves to state 1 with probability 0.3 and a reward of 0.5 and to state UG with probability 0.7 and a reward of 0. Assume no discounting, i.e., $\gamma = 1$. Further, since DG and UG are goal states, the process does not come out of these states once it gets inside any of them and moreover, $J^*(\text{DG}) = J^*(\text{UG}) = 0$.

   (a) Write down the Bellman equations for this problem? (1)

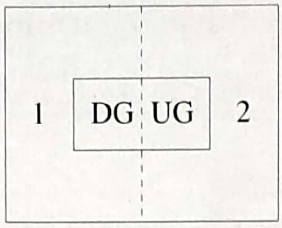

Figure 1: Room Transition Model

(b) Starting with initial value estimates of 0 for both the states, find the value estimates at $n = 3$ using the value iteration procedure, i.e., find $J_3(1)$ and $J_3(2)$, respectively. (2)

4. Consider a mapping $W : \mathcal{R}^{|S|} \to \mathcal{R}^{|S|}$ that is not a contraction but is such that $W^p : \mathcal{R}^{|S|} \to \mathcal{R}^{|S|}$ is a contraction for some $p > 1$. Here W^p is the composition of W with itself p times.

(a) Show that there exists a unique $U^* \in \mathcal{R}^{|S|}$ such that $U^* = WU^*$. (3)

(b) Show that for any $U \in \mathcal{R}^{|S|}$, $\lim_{n\to\infty} W^n U = U^*$. (3)