CS6700: Reinforcement Learning

Homework 1: **Due on September 2, 2018, Max marks: 30**

1. Consider a finite horizon MDP with N stages. Suppose there n possible states in each stage and m actions in each state. Why is the DP algorithm computationally less intensive as compared to an approach that calculates the expected cost J^π for each policy π? Argue using the number of operations required for both algorithms, as a function of m, n and N. (3 marks)

2. Consider the finite horizon MDP setting, as formulated in Section 1.2 of the course notes. In place of the expected cost objective defined there, consider the following alternative cost objective for any policy π and initial state x_0:

$$J_\pi(x_0) = \mathbb{E}\left[\exp\left(g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), x_{k+1}]\right)\right],$$

Answer the following: (4+2 marks)

   (a) Show that an optimal cost and an optimal policy can be obtained by the following DP-algorithm variant:

$$J_N(x_N) = \exp\left(g_N(x_N)\right),$$
$$J_k(x_k) = \min_{a_k \in A(x_k)} \mathbb{E}_{x_{k+1}}\left(\exp\left(g_k(x_k, a_k, x_{k+1})\right) J_{k+1}(x_{k+1})\right).$$

   (b) Let $V_k(x_k) = \log J_k(x_k)$. Assume that the single stage cost g_k is a function of x_k and a_k only (and does not depend on x_{k+1}). Then, show that the DP algorithm, which is specified above, can be re-written as

$$V_N(x_N) = g_N(x_N),$$
$$v_k(x_k) = \min_{a_k \in A(x_k)} \left(g_k(x_k, a_k) + \log \mathbb{E}_{x_{k+1}}\left(\exp\left(V_{k+1}(x_{k+1})\right)\right)\right).$$

3. You are walking along a line of N stores in a shopping complex, looking to buy food before entering a movie hall at the end of the store line. Each store along the line has a probabi1ity p of providing the food you like. You cannot see what the next store (say $k+1$) offers, while you are at the kth store and once you pass store k, you cannot return to it. You can choose to buy at store k, if it has the food item you like and pay an amount $N-k$ (since you have to carry this item for a distance proportional to $N-k$). If you pass through all the stores without buying, then you have to pay $\frac{1}{1-p}$ at the entrance to the movie hall to get some food.

   Answer the following: (2+2+4 marks)

   (a) Formulate this problem as a finite horizon MDP.

   (b) Write a DP algorithm for solving the problem.

   (c) Characterize the optimal policy as best as you can. This may be done with or without the DP algorithm.

Hint:

- *Follow the technique from the asset selling example, since the question is about when to stop.*
- *Argue that if it is optimal to buy at store k when the item you like is available, then it is optimal to buy at store $k+1$ when the likeable food is available there.*
- *Show that $\mathbb{E}(J_{k+1}(x))$ is a constant that depends on $n-k$, leading to an optimal threshold-based policy.*

4. Suppose there are N jobs to schedule on a computer. Let T_i be the time it takes for job i to complete. Here T_i is a positive scalar. When job i is scheduled, with probability p_i a portion β_i (a positive scalar) of its execution time T_i is completed and with probability $(1-p_i)$, the computer crashes (not allowing any more job runs). Find the optimal schedule for the jobs, so that the total proportion of jobs completed is maximal.

   Hint: Use an interchange argument to show that it is optimal to schedule jobs in the order $\dfrac{p_i\beta_i Z_i}{1-p_i}$, where Z_i is the residual execution time of job $i \in \{1,\ldots,N\}$. (4 marks)

5. Consider a finite horizon MDP setting, where the single stage cost is time-invariant, i.e., $g_k \equiv g$, $\forall k$. Also, assume that the same set of states and actions are available in each stage.

   Answer the following: (2+2 marks)

   (a) If $J_{N-1}(x) \leq J_N(x)$ for all $x \in \mathcal{X}$, then,
   $$J_k(x) \leq J_{k+1}(x), \text{ for all } x \in \mathcal{X}, \text{ and } k.$$

   (b) If $J_{N-1}(x) \geq J_N(x)$ for all $x \in \mathcal{X}$, then,
   $$J_k(x) \geq J_{k+1}(x), \text{ for all } x \in \mathcal{X}, \text{ and } k.$$

6. Suppose that the course notes for CS6700 has been written up and the number of errors in this document is X (a non-negative scalar). Assume that there are N students willing to proofread the notes. The proofreading exercise proceeds in a serial fashion, with each student charging the same amount, say $c_1 > 0$. The kth student can find an error in the notes with probabi1ity p_k, and this is independent of the number of errors found by previous students. The course instructor has an option to stop proofreading and publish the notes or continue proofreading. Each undetected error in the published notes costs $c_2 > 0$.

   Answer the following: (1+2+4+2 marks)

   (a) Formulate this problem as a finite horizon MDP.

   (b) Write a DP algorithm for solving the problem.

   (c) Analyze the DP algorithm to characterize as best as you can the optimal selling policy.

   (d) Discuss the modifications to the DP algorithm for the case when the number of errors X is a r.v.