CS6700: Reinforcement learning

Quiz - 1

Course Instructor : Prashanth L.A.

Date : Sep-8, 2018 **Duration** : 120 minutes **Max marks** : 30

INSTRUCTIONS: Answers should be given with proper justification. Please use rough sheets for any calculations *if necessary.* Please **DO NOT** submit the rough sheets. Please DO NOT use pencil for writing the answers.

Assume standard data whenever you feel that the given data is insufficient. However, please do quote your assumptions explicitly.

1. True or False? (Answer any six) (9 marks)

   Note: 1.5 *marks for the correct answer and* -0.5 *for the wrong answer.*

   (a) In an SSP problem setting, suppose that each improper policy π has $J_\pi(i) = \infty$ for at least one state i. Then, there is no improper policy π' such that $J_{\pi'}(j) = -\infty$ for at least one state j.

   (b) In a deterministic shortest path problem, if there are no cycles, then every improper policy π has infinite cost J_π for at least one state.

   (c) Let π be a proper policy. Suppose there exists a stationary policy π' such that $T_{\pi'} J_\pi = T J_\pi$, then π' is proper.

   (d) There exists a policy π such that $\lim_{k\to\infty} T_\pi^k J < \lim_{k\to\infty} T^k J$ for some vector J.

   (e) In a SSP problem, if the single stage cost $g(x, a, x')$ is replaced by $g'(x, a, x') = g(x, a, x') + 10$, $\forall (x, a, x')$, then the optimal policy remains unaffected.

   (f) In a finite horizon MDP setting, suppose we have time-invariant state and action spaces. Consider a modified problem, where the terminal cost $g_N(x)$ is replaced by $g'_N(x) = g_N(x) + 10$. Let J_k and J'_k denote the kth stage functions in the DP algorithm for the original and modified problems. Then, $J_k(x) \leq J'_k(x)$, for all x and k.

   (g) In a SSP problem, if $T_\pi J^* = T J^*$ for a stationary policy π, then π is optimal.

2. A finite horizon MDP setting is as follows:

   Horizon $N = 2$, states $\{1, 2\}$ and actions $\{a, b\}$ (available in each state). Transition probabilities are given by

$$p_{11}(a) = p_{12}(a) = \frac{1}{2}; \quad p_{11}(b) = \frac{1}{4}, p_{12}(b) = \frac{3}{4};$$
$$p_{21}(a) = \frac{2}{3}, p_{22}(a) = \frac{1}{3}; \quad p_{21}(b) = \frac{1}{3}, p_{22}(b) = \frac{2}{3};$$

   The time-invariant single-stage costs are as follows:

$$g(1, a) = 3, g(1, b) = 4, g(2, a) = 2, g(2, b) = 1.$$

   There is no terminal cost. Calculate the optimal expected cost using the DP algorithm and specify an optimal policy. (5 marks)

3. Suppose you want to travel from a start point S to a destination point D in minimum average time. You have the following route options:

   (a) A direct route that requires α time units.

   (b) A potential shortcut that requires β time units to get to a intermediate point I. From I, you can do one of the following: (i) go to D in c time units; or (ii) head back to S in β time units. The random variable c takes one of the values $c_1, \ldots, c_m$, with respective probabilities $p_1, \ldots, p_m$. The value of c changes each time you return to S, independent of the value in the previous time period.

   Answer the following: (5+1+4+5 marks)

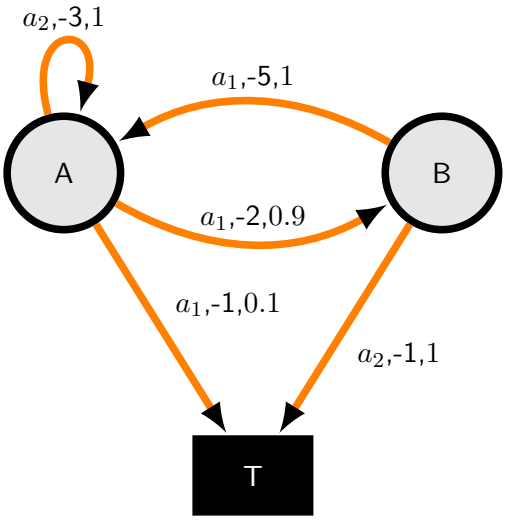
   (a) Formulate the problem as an SSP problem. Write the Bellman's equation ($J = TJ$) and characterize the optimal policy as best as you can.

   (b) Are all policies proper? If not, why does the Bellman equation hold?

   (c) Solve the problem for $\alpha = 2, \beta = 1$, $c_1 = 0, c_2 = 5$, and $p_1 = p_2 = \frac{1}{2}$. Specify the optimal policy.

   (d) Consider the following problem variant, where at I, you have the additional option of waiting for d time units. Each d time units, the value of c changes to one of $c_1, \ldots, c_m$ with probabilities as before, independently of the value at the previous time period. Each time c changes, you have the option of waiting extra d units, return to the start or go to the destination. Write down the Bellman equation and characterize the optimal policy.

4. Consider the MDP specified through the transition diagram below, where the edge labels are in the following format: (action (a), cost $g(i, a, j)$, probability $p_{ij}(a)$).



   Consider the following policies: $\pi_1 = (a_1, a_2)$ (i.e., take action a_1 in state A and a_2 in state B) and $\pi_2 = (a_1, a_1)$. Answer the following: (1+4 marks)

   (a) Are π_1 and π_2 proper?

   (b) Calculate TJ, $T_{\pi_1}J$, and $T_{\pi_2}J$ for $J = [15, 10]$.