CS6700: Reinforcement learning

Quiz - 2

Course Instructor : Prashanth L.A.

Date : Oct-7, 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? (3 marks)

   Note: 1 *marks for the correct answer and* -0.25 *for the wrong answer.*

   (a) Let $\{\pi_k\}$ be the sequence of policies generated by the policy iteration algorithm. Then $J_{\pi_{k+1}} = J_{\pi_k}$ if and only if $J_{\pi_k} = J^*$.

   (b) If $f : \mathbb{R}^n \to \mathbb{R}^n$ has a fixed point, then it is a contraction mapping w.r.t. the euclidean norm.

   (c) Let S, ρ be a metric space, where S is the set, and ρ the metric. If a function $f : S \to S$ is a contraction mapping, then there exists a unique x^* such that $f(x^*) = x^*$.

2. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a contraction mapping with modulus α. We know that there exists an x^* such that $f(x^*) = x^*$. Show that (3 marks)

$$\|x^* - f^n(x)\| \leq \frac{\alpha^n}{1-\alpha} \|f(x) - x\|, \quad \forall x \in \mathbb{R}^n, n \geq 1.$$

3. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a contraction mapping w.r.t. the sup-norm. Let α be the modulus, and x^* the fixed point of f. Suppose f satisfies

$$f(x + ce) = f(x) + \alpha ce, \text{ for all } x \in \mathbb{R}^n \text{ and scalar } c.$$

   In the above, e denotes the n-vector of ones. Assume that there exists a $x' \in \mathbb{R}^n$, and scalar b such that $f(x') - x' \leq b \cdot e$.

   Show that (4 marks)

$$x^* \leq f(x') + \frac{\alpha}{1-\alpha} b \cdot e \leq x' + \frac{1}{1-\alpha} b \cdot e.$$

4. Consider a discounted cost MDP with two states, denoted 1 and 2. In state 1, there are two feasible actions, say a and b, while in state 2, there is only one action available, say c. The transition probabilities are given by

$$p_{11}(a) = p_{12}(a) = 0.5; \quad p_{11}(b) = 0, \ p_{12}(b) = 1;$$
$$p_{21}(c) = 0, \ p_{22}(c) = 1;$$

   The single-stage costs are as follows:

$$g(1, a) = -5, \ g(1, b) = -10; \quad g(2, c) = 1.$$

   Answer the following: (5 + 5 marks)

(a) Use the Bellman equation to calculate the optimal policy, when the discount factor α is set to 0.1, 0.5 and 0.9.

(b) Start with a policy that chooses action b in state 1, and perform policy iteration with $\alpha = 0.95$. Does the algorithm converge in two steps? Show your calculations for each policy iteration step.

5. A manufacturer at each time period receives an order for her product with probability p, and receives no order with probability $1 - p$. At any period, she has a choice of processing all the pending orders in a batch, or process no order at all. The cost per pending order in any period is $c > 0$, and the setup cost to process the pending orders is $K > 0$. She aims to find an optimal processing policy that minimizes, in expectation, the infinite horizon discounted cumulative cost, with a discount factor $\alpha \in (0, 1)$. The setting has a constraint that the maximum number of orders that can be pending is n. Answer the following: $(3 + 6 + 4$ marks$)$

(a) Letting the number of pending orders as the state, write down the Bellman equation for this problem.

(b) Let J^* denote the optimal cost vector. Show that $J^*(i)$ is monotonically non-decreasing in i. Using this fact, show that a threshold-based policy is optimal.

(c) Show that if policy iteration algorithm is initialized with a threshold policy, every subsequently generated policy will be a threshold policy.