Exam Markov Decision Theory and Algorithmic Methods (191531920)

April 11, 2016 8:45-11:45 hrs

This exam consists of 4 exercises.
Motivate all your answers.

1. Consider the following infinite-horizon Markov Decision Problem (MDP) with the average reward criterion. Decisions are taken at times 1, 2, 3, ... There are two states, $S = \{s_1, s_2\}$, with actions $a_{1,1}$ and $a_{1,2}$ in state s_1, and action $a_{2,1}$ in state s_2. Further, $r(s_1, a_{1,1}) = 3$, $r(s_1, a_{1,2}) = 4$, $r(s_2, a_{2,1}) = 5$, and $p(s_1|s_1, a_{1,1}) = 0$, $p(s_1|s_1, a_{1,2}) = 1/2$, $p(s_1|s_2, a_{2,1}) = 3/4$, $p(s_2|s, a) = 1 - p(s_1|s, a)$.

   (a) Use the optimality equations to determine an average optimal policy, and the value of this MDP.

   (b) Why does there exist a stationary average optimal policy?

   (c) Let $\varepsilon > 0$. Is it possible that the value iteration algorithm results in an optimal policy, instead of an ε-optimal policy?

   (d) Suppose g^* and h^* solve the optimality equations. Let d^* be h^*-improving. Prove that $(d^*)^\infty$ is an average optimal policy.

2. Consider the following infinite-horizon MDP with the discounted reward criterion, and discount factor $\lambda = 0.9$. Decisions are taken at times 0, 1, 2, ... There are two states, $S = \{s_1, s_2\}$, with actions $a_{1,1}$ and $a_{1,2}$ in state s_1, and actions $a_{2,1}$ and $a_{2,2}$ in state s_2. Further, $r(s_1, a_{1,1}) = 5$, $r(s_1, a_{1,2}) = 10$, $r(s_2, a_{2,1}) = -1$, $r(s_2, a_{2,2}) = -1$ and $p(s_1|s_1, a_{1,1}) = 1/2$, $p(s_1|s_1, a_{1,2}) = 0$, $p(s_1|s_2, a_{2,1}) = 0$, $p(s_1|s_2, a_{2,2}) = 1$. We solve this MDP with linear programming.

   (a) Formulate the dual LP that corresponds to the given MDP.

   (b) Why do we prefer to solve the dual LP over the primal LP?

   (c) Give $\alpha(s_1) = 0.4$, the optimal solution of the dual LP is x^* with $x^*(s_1, a_{1,1}) = 0$, $x^*(s_1, a_{1,2}) = 4.95$, $x^*(s_2, a_{2,1}) = 0$, and $x^*(s_2, a_{2,2}) = 5.05$. Use this solution to construct an optimal policy of the given MDP.

3. Consider a large-scale MDP with the average reward criterion. Let $S = \{1, 2, 3, \ldots\}$, and $A_s = \{a_{s,1}, a_{s,2}\}$ for all $s \in S$. Further, $r(s, a_{s,1}) = 0$, $r(s, a_{s,2}) = 1 - 1/s$, and $p(s+1|s, a_{s,1}) = 1$, $p(s|s, a_{s,2}) = 1$, and all other transition probabilities are equal to 0.

   (a) Consider the history dependent policy π^* which, for each state $s \in S$, uses action $a_{s,2}$ s times in state s, and then uses action $a_{s,1}$ once. Show that this policy, starting in state 1, generates the reward stream $(0, 0, 1/2, 1/2, 0, 2/3, 2/3, 2/3, 0, 3/4, \ldots)$.

   (b) Determine $g_-^{\pi^*}$ and $g_+^{\pi^*}$ for the reward stream given in part (a).

   (c) You are told that $g^* = 1$. Why does there not exist a deterministic stationary optimal policy in this MDP?

   (d) Consider a general large-scale MDP with the average reward criterion, and finite A_s. Suppose the following assumptions hold.

   - For each $s \in S$, $-\infty < r(s, a) \leq R < \infty$.
   - For each $s \in S$ and $0 \leq \lambda < 1$, $v_\lambda^*(s) > -\infty$.
   - There exists a $K < \infty$ such that for each $s \in S$ $h_\lambda(s) \leq K$ for $0 \leq \lambda < 1$.
   - There exists a nonnegative function $M(s)$ such that
     i. $M(s) < \infty$;
     ii. for each $s \in S$, $h_\lambda(s) \geq -M(s)$ for all λ, $0 \leq \lambda < 1$; and
     iii. for each $s \in S$ and $a \in A_s$, $\sum_{j \in S} p(j|s, a) M(j) < \infty$.

   What can you say about the gain g and bias h?

4. Approximate dynamic programming (adp) is a recent technique, useful for solving large-scale MDPs.

   (a) Describe and explain the basic adp algorithm.

   (b) What is the post-decision state? Mention an advantage and a disadvantage of using this state in adp.

Norm:

1				2			3				4		total
a	b	c	d	a	b	c	a	b	c	d	a	b	
4	2	3	4	3	2	3	2	2	2	3	4	2	$+ 4 = 40$

Exam grade: (Points obtained)/4