Exam Markov Decision Theory and Algorithmic Methods (191531920)

January 25, 2013 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 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 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 $p(s_1|s_1, a_{1,1}) = 0$, $p(s_1|s_1, a_{1,2}) = 1/2$, $p(s_1|s_2, a_{2,1}) = 3/4$.

   (a) The policy maker has to select a decision rule for each decision epoch. Give the definition of a decision rule. Also, describe the four classes of decision rules.

   (b) For a given stationary policy d^∞ the MDP reduces to an MRP. The optimality equations of an MRP are

   $$(I - P)g = 0 \quad \text{and} \quad g + (I - P)h = r.$$

   Prove that if vectors g and h satisfy these equations, then $g = P^*r$ and $h = H_P r + u$ with u such that $(I - P)u = 0$.

   (c) Let $d_1(s_1) = a_{1,2}$ and $d_1(s_2) = a_{2,1}$. Determine the gain g and bias h of the resulting MRP.

   (d) Is the policy $(d_1)^\infty$ average optimal?

   (e) Two ways of solving average reward MDPs are policy iteration and linear programming. Mention one difference and one similarity between these methods.

2. Consider the same Markov decision problem as in exercise 1, only with two changes. First, in state s_2 there is a second action $a_{2,2}$ available with $r(s_2, a_{2,2}) = 6$, and $p(s_1|s_2, a_{2,2}) = 1$. Second, consider the discounted reward criterion with discount factor λ.

   (a) What is the relation between v_λ^* and g^*, the optimal gain of the same MDP with average reward criterion?

   (b) Perform one iteration of the value-iteration algorithm. Use starting value $v^0 = (4, 2)$ and $\varepsilon = 0.1$.

   (c) Let $\{v^n\}$ denote the iterates of value iteration. Do we have monotone convergence (that is, $v^{n+1} \geq (\leq) v^n$ for all n)?

3. Consider a discounted MDP with countable state space $S = \{0, 1, 2, \ldots\}$.

   (a) Give two reasons why the small-scale Markov decision theory is not applicable for such an MDP.

   (b) When dealing with 'unbounded' rewards, we need the following assumptions.

      i. There exists a constant $\mu < \infty$ such that $\sup_{a \in A_s} |r(s,a)| \leq \mu w(s)$.

      ii. There exists a constant κ, $0 \leq \kappa < \infty$, for which $\sum_{j \in S} p(j|s,a) w(j) \leq \kappa w(s)$.

      iii. For each λ, $0 \leq \lambda < 1$, there exists an α, $0 \leq \alpha < 1$, and an integer J such that $\lambda^J \sum_{j \in S} P_\pi^J(j|s) w(j) \leq \alpha w(s)$ for all $\pi = (d_1, \ldots, d_J)$ where $d_k \in D^{MD}$, $k \in \{1, 2, \ldots, J\}$.

   Suppose these hold. What can you say about the solution(s) of the optimality equation?

   (c) Consider the following setting. Let $A_s = \{0, 1, 2, \ldots, M\}$, $r(s,a) = s$, and $p(j|s,a) = 1$ if $j = s + a$ and 0 otherwise. Show that there exists a function w (which one?) such that assumptions i - iii, as stated above, hold.

4. Consider a large-scale MDP with the discounted reward criterion.

   (a) One method used in approximate dynamic programming (adp) is aggregation. Suppose we apply aggregation to our MDP. Describe the transition probabilities in the aggregated system as a function of the transition probabilities of the original system.

   (b) Using aggregation, explain how to obtain a suboptimal policy when the control (action) is applied with knowledge of the aggregate state. Assume the control sets $U(i)$ are independent of the state i.

   (c) Another approach in adp works with Q-factors. How are Q-factor defined? What is their purpose?

   (d) Describe the Q-learning algorithm. Mention two conditions that must be satisfied to guarentee convergence of the Q-learning algorithm. (There are more.)

Norm:

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