Exam Markov Decision Theory
and Algorithmic Methods (191531920)

April 13, 2015 3 hrs

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

1. Consider an infinite horizon average reward Markov Decision Problem (MDP) with state space $S = \{s_1, s_2\}$, and action sets $A_{s_1} = \{a_{1,1}, a_{1,2}\}$, $A_{s_2} = \{a_{2,1}, a_{2,2}\}$. The immediate rewards are $r(s_1, a_{1,1}) = 4$, $r(s_1, a_{1,2}) = 6$, $r(s_2, a_{2,1}) = -4$, $r(s_2, a_{2,2}) = -6$. The transition probabilities are given by $p(s_2|s_1, a_{1,1}) = 1/2$, $p(s_2|s_1, a_{1,2}) = 1$, $p(s_2|s_2, a_{2,1}) = 1/2$, and $p(s_2|s_2, a_{2,2}) = 0$.

   (a) The stationary policy d^∞ is defined by the decision rule d satisfying $d(s_1) = a_{1,2}$ and $d(s_2) = a_{2,2}$. Calculate the gain g^{d^∞} of this stationary policy.

   (b) The optimality equations in vector notation are given by $B(g, h) = 0$ with

   $$B(g,h)(s) = \max_{a \in A_s} \left\{ r(s,a) - g + \sum_{j \in S} p(j|s,a)h(j) - h(s) \right\}.$$

   Write down the optimality equations for this MDP. Use these equations and their properties to show that the optimal gain g^* is bounded, namely $-4 \leq g^*(s) \leq 6$.

   (c) Show that $g^*(s) = 2/3$, $s \in S$, is the optimal gain.

   (d) Let $h = (10/3, -10/3)$ be a bias vector. Determine a decision rule d that is h-improving.

   (e) Suppose that you are asked to check whether or not a given policy π is average optimal. Mention two different ways to do so.

2. (a) Consider an infinite horizon discounted MDP. Explain in words why we may restrict attention to Markov policies, instead of history-dependent policies, when analyzing discounted MDPs.

   (b) One algorithm for solving infinite horizon discounted MDPs is the value iteration algorithm. This results in a stationary policy $(d_\varepsilon)^\infty$ with

   $$d_\varepsilon(s) \in \arg\max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j|s,a) v^{n+1}(j) \right\}$$

   for each state $s \in S$. Prove that the policy $(d_\varepsilon)^\infty$ is ε-optimal.

3. Consider large-scale MDPs with countable state space $S = \{0, 1, \ldots\}$, discount factor λ, and unbounded rewards.

   (a) In this setting, the weighted supremum norm with respect to w is used: $\|v\|_w = \sup_{s \in S} w(s)^{-1}|v(s)|$ with w an arbitrary positive real-valued function on S satisfying $\inf_{s \in S} w(s) > 0$.

   Let $A_s = \{0, 1, 2, \ldots, M\}$ for all states s, $r(s, a) = s$, and $p(j|s, a) = 1$ if $j = s + a$ and $p(j|s, a) = 0$ else. Let $w(s) = \max(s, 1)$. Show that there exists a constant κ, $0 \leq \kappa < \infty$, such that

   $$\sum_{j \in S} p(j|s, a)w(j) \leq \kappa w(s), \qquad \text{for all } a \in A_s, \text{ for all } s \in S.$$

   (This is one of the conditions for existence of an optimal policy.)

   (b) Under suitable conditions (one of them is mentioned in part (a)), the optimality equation has an optimal solution; that is, the MDP has a value. Why do algorithms like the value iteration algorithm not work in this case? And, how can we approximate the value in practice?

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

   (a) Describe and explain the basic adp algorithm.

   (b) Describe and explain the Q-learning algorithm. What is the advantage of using this algorithm? How is it related to exploration?

Points:

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