Exam Markov Decision Theory and Algorithmic Methods (191531920)

January 27, 2017 8:45-11:45 hrs

This exam consists of 4 exercises.
You may bring your own 'cheat sheet' (1 page A4, one-sided).
Motivate all your answers.

1. Consider the following infinite-horizon Markov decision problem (MDP) with the average reward criterion. There are two states, $S = \{s_1, s_2\}$, with actions a and b in state s_1, and action c in state s_2. The rewards are $r(s_1, a) = 2$, $r(s_1, b) = 3$, and $r(s_2, c) = -1$. The transition probabilities are given by $p(s_1|s_1, a) = 1/2$, $p(s_1|s_1, b) = 0$, and $p(s_1|s_2, c) = 1/2$.

   (a) What are the optimality equations for this particular MDP in component notation?

   (b) What is the dual LP corresponding to this MDP? (You do not need to solve this.)

   (c) The optimal solution of the dual LP is $x^*(s_1, a) = 1/2$, $x^*(s_1, b) = 0$, and $x^*(s_2, c) = 1/2$. Use this to construct an average optimal policy.

   (d) How can you use the policy iteration algorithm to check if the policy you found in part (c) is indeed optimal? (You do not have to run this algorithm.)

   (e) Determine the gain g and bias h of this MDP.

2. Consider a finite horizon MDP with horizon 1. Prove that for any Markovian randomized policy in Π^{MR} there is a deterministic policy in Π^{MD} with a reward at least as great.

3. In the book of Puterman it is shown that under mild conditions there exists a value for discounted MDPs with infinite, countable state space. To approximate this value v^*_λ, the finite-state approximation method may be used.

   (a) How can you derive the N-state approximation $v^{N,u}_*$?

   (b) In which two cases does the N-state approximation $v^{N,u}_*$ converge to the value v^*_λ of the MDP?

Please turn over.

4. Consider a finite horizon Markov reward process with the discounted reward criterion. There is a single state, the horizon length is $T = 30$, and the discount factor is $\lambda = 0.95$. The expected discounted reward is

$$F^T = \mathbb{E} \sum_{t=0}^{T} \lambda^t R_t,$$

where $\mathbb{E}$ denotes the expectation. The random variable R_t follows a normal distribution with mean 250 and variance 150, and is assumed to be independent of prior history. In this simple setting we use approximate dynamic programming to estimate F^T.

(a) What is the exact value of F^T?

(b) Let $\hat{v}_t^n = \sum_{t'=t}^{T} \lambda^{t'-t} R_{t'}(\omega^n)$, where ω^n represents the nth sample path, and $R_{t'}^n = R_{t'}(\omega^n)$ is the realization of the random variable $R_{t'}$ for the nth sample path. Show that $\hat{v}_t^n = R_t^n + \lambda \hat{v}_{t+1}^n$. What is the meaning of $\hat{v}_0^n$?

(c) Formulate an adp algorithm to estimate F^T. In the value function updating equation use the stepsize $\alpha_{n-1} = 1/n$ for iteration n.

(d) The stepsize is used for smoothing. What is the goal of smoothing, and why is it needed?

Points:

1	2	3	4	total
16	4	6	10	+ 4 = 40

Exam grade: (Points obtained)/4