CS286r Multi-Agent Learning and Implementation
Homework 3: Reinforcement Learning and MDPs

Spring Term 2006
Prof. David Parkes
Division of Engineering and Applied Sciences
Harvard University

Feb 27, 2006

Due: Monday 3/6/2006, at the beginning of class. You may use any sources that you want, but you must cite the sources that you use. You can also work in a group, just list off the people you're working with. *Work hard on making the proofs clear, concise, and easy to read.* Total points: 100

1. (20 pts) Devise three example tasks of your own that fit into the reinforcement learning framework, identifying for each its states, actions and rewards. Make the three examples as *different* as possible.

2. (20 pts) Consider the *gridworld* example on p.78–79 of the examples handout and introduced on p.71.[1]

   (a) Establish that the Bellman equation, $V^*(s) = \max_{a \in A} \left(r(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V^*(s') \right)$, holds for the optimal policy in the middle state.

   (b) Explain why the policy is as illustrated in the RHS of Fig 3.8, given V^*.

   (c) Provide some intuition for why this policy is optimal.

3. (20 pts) The *policy improvement theorem* argues that for any pair of deterministic policies, π and π', such that for all $s \in S$, $Q^\pi(s, \pi'(s)) \geq V^\pi(s)$, then policy π' is at least as good as π, with $V^{\pi'}(s) \geq V^\pi(s)$ for all $s \in S$. Prove this result and explain why it establishes the monotonic improvement property asserted for policy improvement on p.9 of the lecture notes.

4. (20 pts) (a) Why is Q-learning considered an *off policy* learning method?

   (b) Consider a learning algorithm that is modified from Q-learning, with update-rule

   $$Q(s_t, a_t) := (1-\alpha) Q(s_t, a_t) + \alpha \left(r(s_t, a_t) + \gamma Q(s_{t+1}, \pi(s_{t+1})) \right)$$

   in place of the standard rule. Explain how the rule is different, and whether this new method an on-policy or off-policy method?

   (c) Given the same amount of experience, would you expect this method to work better or worse than SARSA?

5. (20 pts) An *optimal stopping problem* is a special case of an MDP in which there is only one action available in each state: *continue* or *stop*. Formulate this problem as a stopping problem:

   When going out for dinner, I always try to park as close as possible to the restaurant. I do not like to pay to use a parking lot, so I always seek on-street parking. My chosen restaurant is on a very long street running east to west which allows parking on one side only. I approach the restaurant from the east starting at a distance of Q units away. Because traffic is heavy I can only check one potential parking spot at a time. In your formulation, assume the street is divided into one-car-length sections, and the probability that a parking spot at distance s from the restaurant is unoccupied is p_s, and independent of all others.

6. Formulate one of your problems in Q#1 as an MDP.

[1] Look online here: `http://www.eecs.harvard.edu/~parkes/cs286r/spring06/lectures/rlsuttonbarto.pdf`.