CS 294-67 Reinforcement Learning
Spring 2011 Stuart Russell
Assignment 1

Due 3/9/11

1. Prove invariance of optimal policies under the affine transformation

   $$R'(s,a,s') = c \cdot R(s,a,s') + d \text{ for } c > 0 \ .$$

2. In class we proved that optimal policies are invariant under the potential-based transformation

   $$R'(s,a,s') = R(s,a,s') + \gamma\Phi(s') - \Phi(s).$$

   What is the most general optimality-preserving transformation obtainable by combining affine and potential-based transformations? Prove your claim.

3. Derive an incremental-update equation for computing a running sample variance, analogous to the incremental computation of running averages.

4. Is it possible to apply Monte Carlo policy evaluation methods to environments with discounted rewards and no terminal states? How might this work and what problems might arise?

5. The *stochastic-arrival vacuum world* is a family of MDPs with k edge-connected squares, arbitrarily arranged. Each square can be clean or dirty; on any given time step, the probability that any given clean square becomes dirty is p. The actions are $Suck$, $NoOp$, $Left$, $Right$, Up, and $Down$, all with predictable effects. The cost of sucking is 1, the cost of moving is 2, and there is a penalty of 1 per time step for each dirty square.

   (a) How many states are there?

   (b) Suppose we formulate the problem as an episodic problem where the terminal state has no dirt. Is there a proper policy? Does every improper policy have value $-\infty$ for some state?

   (c) Using a DP algorithm of your choice, devise optimal (or ϵ-optimal) policies for the largest environment you can handle, for different values of γ and p.

   (d) Examine and comment on the resulting behaviors. Does your agent ever decide to do nothing? Why (not)?

   [Note: this question is deliberately open-ended—I am leaving it up to you to decide what experiments to run and where to look for interesting and possible explicable phenomena. You are welcome to do this part in pairs.]

6. What might be a good function approximator for very large stochastic-arrival vacuum worlds?

7. Discuss the feasibility of applying Q-learning to (a) a continuous version of the stochastic-arrival vacuum world, and (b) a partially observable version in which dirt sensing is limited to the current square.