CS 294-7 Reinforcement Learning
Spring 2003 Stuart Russell Assignment 1

Due 2/26/03

1. Let V be a value function with max-norm error ϵ and let policy π be greedy w.r.t. V. Show that

   $$||V^\pi - V^*|| < 2\epsilon\gamma/(1-\gamma) \ .$$

2. Starting from the Bellman equation for Q^*, define Q-iteration (analogous to value iteration) and prove that it converges to the optimal Q-function in the discounted case.

3. Consider the following Monte Carlo policy evaluation method:

   - Choose in advance an integer $K(s)$ for each state s.
   - Run policy π for enough complete episodes such that each state s has been visited $K(s)$ times.
   - For each s, estimate $\hat{V}^\pi(s)$ by $\frac{1}{K}\sum_{k=1}^{K} S(s,k)$ where $S(s,k)$ is the total reward received from the kth visit to state s to the end of the corresponding episode.

   Calculate $E[\hat{V}^\pi(s)]$ for the case $K(1)=2$ in a two-state episodic MDP with $T(1,a,0)=p$, $T(1,a,1)=1-p$, and $R(1,a,0)=R(1,a,1)=1$. Is $\hat{V}^\pi(s)$ biased in this case? If so, explain why. If not, explain how it can be unbiased given that the rewards after the first visit to s in each episode are always greater than the rewards after subsequent visits in the same episode, and given that the first-visit estimator is unbiased.

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 policies for the largest environment you can handle, for the episodic problem and for the discounted problem with different values of γ. Does the runtime depend on the value of 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.]

6. Consider the *partially observable* stochastic arrival vacuum world in which the agent can sense only if its *current* location is dirty. Can you say anything about what the optimal policy might be?