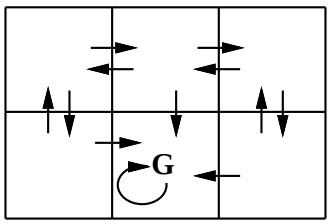
15-681: Machine Learning

Homework 6: Reinforcement learning

Due: 8 December 1994

1. Give the value of Q for every state-action pair for the grid-world scenario shown below, assuming $\gamma = 0.9$ and an immediate reward function $R(s, a)$ whose value is 10 for any action leading from an immediate neighbor of G into G (but zero for the action that keeps it in state G). Here the arrows indicate possible actions for the agent. Also indicate the value of $V^*(s)$ and give an optimal policy.



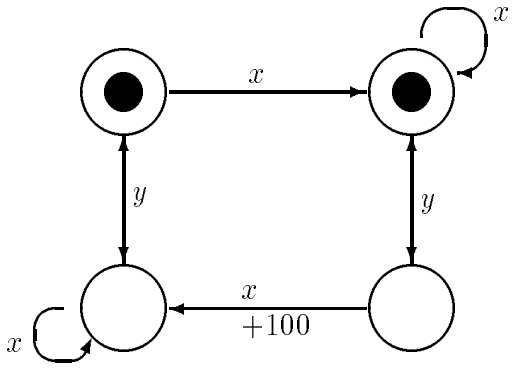
reward 10 for entering G from outside

   Suggest a change to the reward function $R(s, a)$ that alters the $Q(s, a)$ values, but does not alter the optimal policy. Suggest a change to $R(s, a)$ that alters $Q(s, a)$ but that does not affect $V^*(s)$.

2. Explain the effect of initializing Q values before training to highly positive, highly negative, or zero values.

3. Q learning acquires an evaluation function defined over state-action pairs. Why not learn an evaluation function defined over states alone? Consider the function $V(s)$ which assigns to any state s the discounted cumulative reward achieved by executing the optimal policy beginning at state s. Can you derive a V learning algorithm analogous to Q learning? If so, give it. If not, explain the difficulty. (see Sutton's paper on the Adaptive Heuristic Critic for one algorithm to learn V).

4. **[incomplete sensations]** One way to handle sensations that don't determine state is to model the space. Another is to remember your past actions.

   Show that the following deterministic environment has no optimal policy in which the action is a function of the current sensation, but *does* have one where the action is a function of the current sensation and the previous action. For our purposes, lets say that the goal is to maximize *average* reward, and that a policy is *optimal* if for *any* starting configuration (sensation $\times$ memory), after a finite number of steps it gets into a configuration where it is now doing best possible (over all unlimited policies).

(Actions are x, y.
4 states, 2 sensations.
One edge has reward +100
and the rest are 0.)