Markov Decision Processes & Reinforcement Learning

1. **Value Iteration**

   Say an MDP has state space S and reward R where all rewards are positive. If you run value iteration, what is the largest k for which $V_k(s)$ is zero?

2. **Infinite Horizon**

   You are on a linear space and can move only right or left. Each position has reward r_i and $\gamma \approx 1$. Describe your optimal policy given any state i (don't forget about ties).

3. **Value Iteration vs Expectimax Search**

What is the running time of Expectimax Search and Value Iteration as a function of the horizon, T, the number of actions, M, the the number of transitions for any state and action, L, and the number of steps, N? Why don't we always choose the algorithm with better asymptotic running time?

4. **Setting Rewards to Compute Shortest Path**

Suppose we have a grid-world where each state is represented as a point in $\{(x, y) | 0 \leq x \leq 4, 0 \leq y \leq 4\}$. Suppose you have a robot that starts in the lower left corner and is given a goal point that is needs to reach. At each state, the robot can move one point to the left, right, up, or down, and each action has a 90% chance of success (otherwise you stay at the current point).

- What is the size of the state space in the resulting MDP?
- Suppose we want to minimize the length of path to the goal, but we have no preference for which path the robot should take (and all points are reachable). What rewards could we assign to each state so that we recover a shortest path to the goal as our optimal policy?

5. **Q learning**

Suppose you are standing on a linear board: you can take action L or R (walk left or right). If you walk, you have probability p_a that you actually walk to the next square, where $a \in L, R$. Otherwise, your cat distracted you and you are still on the same square. Staying a square gives you reward r_i. Your learning rate is $\alpha = .5$ and $\gamma = .5$.

(a) You are on square 1, you choose $a = L$ and receive $r = 4$. What is your updated value of $Q(1, L)$?

(b) In the next step, you are on square 0, you choose $a = R$, receive $r = 3$ and end up in $s = 1$. What is your updated value of $Q(0, R)$?

6. **ϵ-Greedy**

Why do we generally use an ϵ-Greedy algorithm when choosing the current action during the Q-Learning algorithm? Describe how you would change the value of ϵ over time to encourage early exploration/learning and good decision making after the state space is sufficiently explored.

7. **Convergence of Q-Learning**

Suppose we have a deterministic world where each state-action pair (s, a) is visited infinitely often. Consider an interval during which every such pair is visited. Suppose that the largest error in the approximation of $\hat{Q}_n$ after n iterations is e_n, i.e.

$$e_n = \max_{s,a} |\hat{Q}_n(s, a) - Q(s, a)|$$

Show that e_{n+1} is bounded above by γe_n, where γ is the usual discount factor.