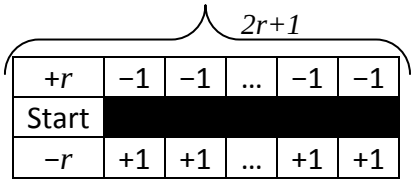
**EECS 391: Introduction to AI (Spring 2012) Written Homework 7 (Max Points: 100)**

**Assigned Friday April 13, due 5pm Tuesday April 24. Write your answers neatly and remember to show all relevant work. Before turning in your work, staple your answer sheets together and write your name(s) and Case ID(s) on the front page.**

1. Sometimes MDPs are formulated with a reward function $R(s,a,s')$ where the reward also depends on the outcome state $s'$. (i) Write the Bellman equation for this scenario. (ii) Show that an MDP with rewards specified as $R(s,a,s')$ can be transformed into another MDP with reward $R(s,a)$ in such a way that it is possible to obtain the optimal policy of the original MDP by solving the new MDP. (15 points)

2. Consider a *two-player* MDP for a zero-sum, turn-based game similar to those we studied earlier. Suppose the players are $A$ and $B$. When the game reaches state $s$, the reward for $A$ is $R(s)$ and the reward for $B$ is $-R(s)$. (i) Let $V_A(s)$ be the value of $s$ when it is $A$'s turn to move, and similarly for $V_B(s)$. (i) Write down the Bellman equations for $V_A(s)$ and $V_B(s)$. (ii) Describe a way to do two-player value iteration with these expressions. (15 points)

3. Consider the $(2r+1)$-by-3 world below. At Start the agent has two deterministic actions, Up and Down. In every other state, the agent has one deterministic action, Right. The states at the right end are terminal states. In each cell a reward for entering that that cell is shown. Start has zero reward. (The omitted cells are identical to their neighbors.) (i) Suppose $r$=50. For what values of $\gamma$ will the agent choose Up at Start? (ii) Suppose $\gamma$=0.75. For what values of $r$ will the agent choose Up at Start? (10 points)




4. Suppose an MDP with $\gamma$=1 has three states, $s_1$, $s_2$ and $s_3$. The agent receives a reward of −1 for entering $s_1$, −2 for $s_2$ and 0 for $s_3$. There are two actions $A$ and $B$ in $s_1$ and $s_2$, and $s_3$ is a terminal state. Action $A$ in $s_1$ moves to $s_2$ with probability 0.8 and stays in $s_1$ with probability 0.2. In $s_2$ it moves to $s_1$ with probability 0.8 and stays in $s_2$ with probability 0.2. In either $s_1$ or s2, action $B$ moves to $s_3$ with probability 0.1 and stays where it is with probability 0.9. (i) Using your intuition, describe the optimal policy in this MDP. (ii) Suppose the initial policy does $B$ in $s_1$ and $s_2$. Apply policy iteration to determine the optimal policy. Does this match your intuition? (iii) Suppose the initial policy does $A$ in $s_1$ and $s_2$. What happens if you apply policy iteration? Why? (15 points)

5. Design suitable features for reinforcement learning in stochastic grid worlds that contain multiple obstacles and multiple terminal states with rewards of +1 or -1. (10 points)

6. Let an approximate value function in an MDP be defined by $V(s)=w_0+w_1x+w_2y+w_3\times sqrt((x-x_0)^2+(y-y_0)^2)$ where $x$, $y$, $x_0$ and $y_0$ are state features. Derive the weight updates for temporal difference learning with this function. (10 points)

7. Suppose $\gamma$=1 and deterministic movement actions. Calculate the true optimal value function and the best linear approximation in $x$ and $y$ for the following grid worlds: (a) a 5-by-5 world with a single +1 reward at the terminal state (5,5) and zero everywhere else. (b) As in (a) but with an additional −1 reward at (5,1). (c) A 5-by-5 world with a single +1 reward at the terminal state (3,3) and zero everywhere else. (15 points)

8. Discuss the similarities and differences between reinforcement learning and Darwinian evolution. Could (model free) reinforcement learning serve as a mathematical model of the evolutionary process? (10 points)