CS188 Fall 2017 Section 5: MDPs and RL

1 MDPs: Micro-Blackjack

In micro-blackjack, you repeatedly draw a card (with replacement) that is equally likely to be a 2, 3, or 4. You can either Draw or Stop if the total score of the cards you have drawn is less than 6. If your total score is 6 or higher, the game ends, and you receive a utility of 0. When you Stop, your utility is equal to your total score (up to 5), and the game ends. When you Draw, you receive no utility. There is no discount ($\gamma = 1$). Let's formulate this problem as an MDP with the following states: $0, 2, 3, 4, 5$ and a *Done* state, for when the game ends.

1. What is the transition function and the reward function for this MDP?

2. Fill in the following table of value iteration values for the first 4 iterations.

States	0	2	3	4	5
V_0					
V_1					
V_2					
V_3					
V_4					

3. You should have noticed that value iteration converged above. What is the optimal policy for the MDP?

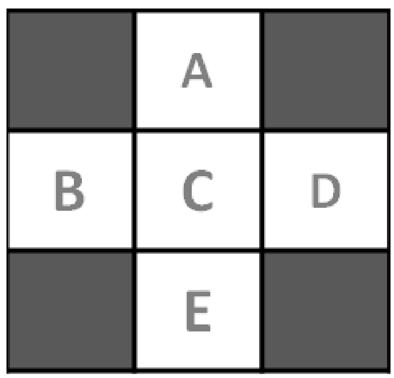
States	0	2	3	4	5
π^*					

4. Perform one iteration of policy iteration for one step of this MDP, starting from the fixed policy below:

States	0	2	3	4	5
π_i	Draw	Stop	Draw	Stop	Draw
V^{π_i}					
π_{i+1}					

2 Learning in Gridworld

Consider the example gridworld that we looked at in lecture. We would like to use TD learning and q-learning to find the values of these states.



Suppose that we have the following observed transitions:
(B, East, C, 2), (C, South, E, 4), (C, East, A, 6), (B, East, C, 2)

The initial value of each state is 0. Assume that $\gamma = 1$ and $\alpha = 0.5$.

1. What are the learned values from TD learning after all four observations?

2. What are the learned Q-values from Q-learning after all four observations?