CS 188
Spring 2025

Regular Discussion 4

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.

(a) What is the transition function and the reward function for this MDP?

(b) 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					

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

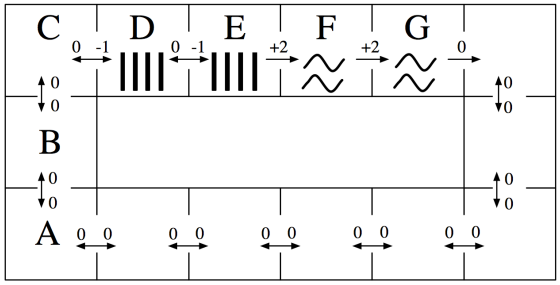
States	0	2	3	4	5
π^*					

(d) 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 MDPs: Grid-World Water Park

Consider the MDP drawn below. The state space consists of all squares in a grid-world water park. There is a single waterslide that is composed of two ladder squares and two slide squares (marked with vertical bars and squiggly lines respectively). An agent in this water park can move from any square to any neighboring square, unless the current square is a slide in which case it must move forward one square along the slide. The actions are denoted by arrows between squares on the map and all deterministically move the agent in the given direction. The agent cannot stand still: it must move on each time step. Rewards are also shown below: the agent feels great pleasure as it slides down the water slide (+2), a certain amount of discomfort as it climbs the rungs of the ladder (-1), and receives rewards of 0 otherwise. The time horizon is infinite; this MDP goes on forever.



(a) How many (deterministic) policies π are possible for this MDP?

(b) Fill in the blank cells of this table with values that are correct for the corresponding function, discount, and state. *Hint: You should not need to do substantial calculation here.* (Note: $V_t(s)$ is the time-limited value of state s, if our Markov Decision Process were to terminate after t timesteps.)

	γ	$s = A$	$s = E$
$V_3(s)$	1.0		
$V_{10}(s)$	1.0		
$V_{10}(s)$	0.1		
$Q_1(s, \text{west})$	1.0	——	
$Q_{10}(s, \text{west})$	1.0	——	
$V^*(s)$	1.0		
$V^*(s)$	0.1		

(c) **(Out of scope until next week. Consider this a sneak peek)** Fill in the blank cells of this table with the Q-values that result from applying the Q-update for the transition specified on each row. You may leave Q-values that are unaffected by the current update blank. Use discount $\gamma = 1.0$ and learning rate $\alpha = 0.5$. Assume all Q-values are initialized to 0. (Note: the specified transitions would not arise from a single episode.)

	$Q(D, \text{west})$	$Q(D, \text{east})$	$Q(E, \text{west})$	$Q(E, \text{east})$
Initial:	0	0	0	0
Transition 1: $(s = D, a = \text{east}, r = -1, s' = E)$				
Transition 2: $(s = E, a = \text{east}, r = +2, s' = F)$				
Transition 3: $(s = E, a = \text{west}, r = 0, s' = D)$				
Transition 4: $(s = D, a = \text{east}, r = -1, s' = E)$				