







## Planning Under Uncertainty - 6 points

Welcome to the world of Game of Thrones! Arya Stark is planning to leave her hometown for an adventure, and your task is to help her enjoy the journey. The map is shown below.

She starts at the location (A, 1). There are two absorbing location states where once she enters them, she cannot leave: (A, 3) stands her enemy “*The Mountain*”, while (E, 5) is the House of Black and White where she has the chance to be trained as a Faceless Man. Two lakes are located in (D, 2) and (E, 3), and Arya cannot pass them. If Arya meets Tywin at (C, 4), she will get a reward of 10. Moreover, if Arya visits either of (E, 1) and (D, 4), she will get a punishment (i.e. a negative reward). We all know that the adventure is usually exhausting, so Arya would get a reward of -1 every time she moves to a blank state.

	A	B	C	D	E
1	START				-5
2					
3	The Mountain  -100				  
4			Tywin  +10	-20	
5					the House of Black and White +100

Suppose Arya can move in singular cardinal directions, i.e. left, right, up, and down. Ties in choosing a direction to go are broken in the order: right > down > left > up. When you select a direction to move in, there is a 70% chance of Arya following your command, and a 15% chance of moving either to her immediate left or right. If an action would move Arya off the map or into the impassable lakes, you can assume that she would stay in her current location. Below are two examples:

- If Arya is at (B, 2) and the command is moving down, there is a 70% chance of her moving down to (B, 3), a 15% chance of moving to her left cell (C, 2), and a 15% chance of moving right to (A, 2).

- If Arya is at (D, 3) and you tell her to move up, she will have a  $70\% + 15\% = 85\%$  (70% moving up; 15% bouncing back after attempting to move right) chance of staying at (D, 3), and a 15% probability of moving left to (C, 3).

The discount factor is 0.8. And the utility (i.e. value) of each location state in the beginning of the adventure (i.e. at time step 0) is shown below:

	A	B	C	D	E
1	0	0	0	0	0
2	0	0	0	x	0
3	3	0	-2	0	x
4	0	-20	0	0	0
5	0	0	0	0	0

## Question 1

**[4 points]** Compute the utility for the following states at time step 1 and 2 using Bellman equation. Some utility values are provided for your reference.

	time step 1	time step 2
(B, 3)	-0.88	-1.8
(C, 3)	-1	4.24
(C, 5)	-1	4.2256
(D, 3)	-1	-1.8
(D, 4)	-20	-14.7744

### Keys:

1. Follow the Bellman Equation in lecture:

$$V(s) = \left[ \max_a \gamma \sum_{s'} P(s'|s,a) V(s') \right] + R(s)$$

2. Utility V.S. Reward
3. Examples of (B, 3) at step 1:

Reward Map :

	A	B	C	D	E
1	-1	-1	-1	-1	-5
2	-1	-1	-1	X	-1
3	-100	-1	-1	-1	X
4	-1	-1	+10	-20	-1
5	-1	-1	-1	-1	+100

	A	B	C	D	E
1	0	0	0	0	0
2	0	0	0	x	0
3	3	0	-2	0	x
4	0	-20	0	0	0
5	0	0	0	0	0

At (B,3):

$$\text{UP: } -1 + 0.8 * (0.7*0 + 0.15*(-2) + 0.15*3) = -0.88$$

$$\text{RIGHT: } -1 + 0.8 * (0.7*(-2) + 0.15*0 + 0 *(-20)) = -2.12$$

## Question 2

**[1 point]** Suppose only for this question that the initial utility for all the location states are zero, all the movements are deterministic, and that the reward map is shown below:

	A	B	C	D	E
1	0	0	0	0	0
2	0	0	0	x	0
3	0	0	0	0	x
4	0	0	0	0	0
5	0	10	0	0	100

Recalling that the discount factor must be in the range  $0 \leq \gamma \leq 1$ , for what range of values for  $\gamma$  is the optimal move “Right” in state (C, 5)?

$0.1 < \gamma \leq 1$

## Question 3

**[1 point]** How is policy iteration different from value iteration? Choose all correct answers below:

- ☒ ~~Policy iteration starts from a random policy and alternates between two steps until convergence: policy evaluation and policy improvement.~~
- ☒ ~~Value iteration starts from estimating the value function and then deduces the policy.~~
- ☒ ~~Policy iteration is guaranteed to converge to the optimal values.~~
- ☒ ~~Value iteration is guaranteed to converge to the optimal values.~~