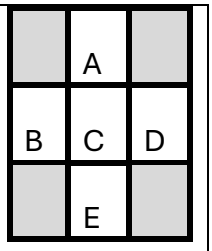
Your Name	
Your UnetID	

Passive Reinforcement Learning

Episode 1: (B,east,C,-1) (C,east,D,-1) (D,exit,x,10)
Episode 2: (B,east,C,-1) (C,east,D,-1) (D,exit,x,10)
Episode 3: (E,north,C,-1) (C,east,D,-1) (D,exit,x,10)
Episode 4: (E,north,C,-1) (C,east,A,-1) (A,exit,x,01)

Assume: $\gamma = 1$



Model-based learning

You can estimate T(s,a,s') and R(s,a,s') by taking the examples you have and weighting them. For example:

$$T(B, east, C) = \frac{2}{2} = 1.00$$

What is your T(C,east,s') transition function? (Hint, it has two s')

What is your R(C,east,D) reward function?

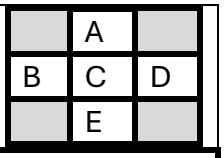
Direct-evaluation

Using the same episodes, you can directly calculate the Utitility values. For example:

$$U(D) = \frac{3}{3}10 \text{ or } \mathrm{U(B)} = \frac{2}{2}(-1 + \gamma * -1 + \gamma^2 * 10) = 8$$

What is the Utility of C?

Assume: $\gamma = 1$, $\alpha = 1/2$
Starting Utilities: U(A)=0, U(B)=0, U(C)=0, U(D)=8, U(E)=0



Temporal-Difference

Consider your first step: (B,east,C,-2). You can calculate a new U(B)

$$U(B) = (1-\alpha) * 0 + \alpha * (-2 + \gamma * 0) = \frac{1}{2} * 0 + \frac{1}{2} * (-2) = -1$$

What is the update with your second step (C,east,D,-2)?

What is the update with your third step (B,east,C,-2) ?