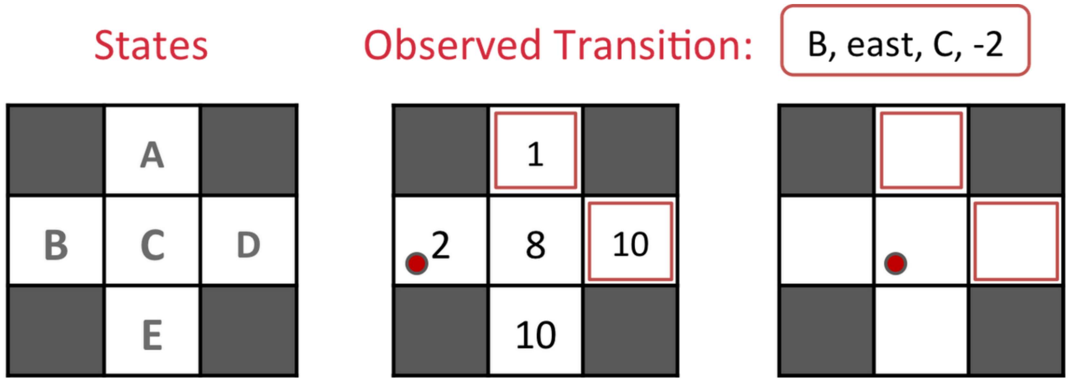
hw5_rl_q4_temporal_difference_learning

Question 4: Temporal Difference Learning

0.0/10.0 points (graded)

Consider the gridworld shown below. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function V^π for each state. A transition is observed, that takes the agent from state B through taking action east into state C, and the agent receives a reward of -2. Assuming $\gamma = 1, \alpha = \frac{1}{2}$, what are the value estimates after the TD learning update? (note: the value will change for one of the states only)



Assume: γ = 1, α = 1/2

$$V^\pi(s) \leftarrow (1-\alpha)V^\pi(s) + \alpha\left[R(s,\pi(s),s') + \gamma V^\pi(s')\right]$$

$\hat{V}^\pi(A) =$

1

Answer: 1

$\hat{V}^\pi(B) =$

4

Answer: 4

$\hat{V}^\pi(C) =$

8

Answer: 8

$\hat{V}^{\pi}(D) =$

10

Answer: 10

$\hat{V}^{\pi}(E) =$

10

Answer: 10

The only value that gets updated is $\hat{V}^{\pi}(B)$, because the only transition observed starts in state B.

$\hat{V}^{\pi}(A) = 1$

$\hat{V}^{\pi}(B) = .5 * 2 + .5 * (-2 + 8) = 4$

$\hat{V}^{\pi}(C) = 8$

$\hat{V}^{\pi}(D) = 10$

$\hat{V}^{\pi}(E) = 10$

Submit