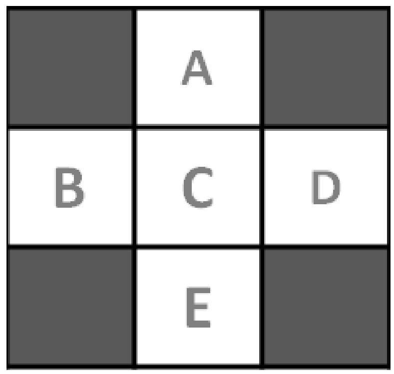
Section 5: RL

1 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?

Q2. Pacman with Feature-Based Q-Learning

We would like to use a Q-learning agent for Pacman, but the size of the state space for a large grid is too massive to hold in memory. To solve this, we will switch to feature-based representation of Pacman's state.

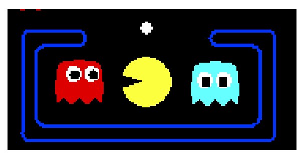
1. We will have two features, F_g and F_p, defined as follows:

$$F_g(s,a) = A(s) + B(s,a) + C(s,a)$$
$$F_p(s,a) = D(s) + 2E(s,a)$$

   where

$$A(s) = \text{number of ghosts within 1 step of state } s$$
$$B(s,a) = \text{number of ghosts Pacman touches after taking action } a \text{ from state } s$$
$$C(s,a) = \text{number of ghosts within 1 step of the state Pacman ends up in after taking action } a$$
$$D(s) = \text{number of food pellets within 1 step of state } s$$
$$E(s,a) = \text{number of food pellets eaten after taking action } a \text{ from state } s$$

   For this pacman board, the ghosts will always be stationary, and the action space is $\{left, right, up, down, stay\}$.



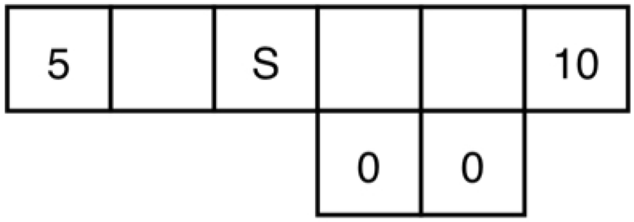
   calculate the features for the actions $\in \{left, right, up, stay\}$

2. After a few episodes of Q-learning, the weights are $w_g = -10$ and $w_p = 100$. Calculate the Q value for each action $\in \{left, right, up, stay\}$ from the current state shown in the figure.

3. We observe a transition that starts from the state above, s, takes action up, ends in state s' (the state with the food pellet above) and receives a reward $R(s,a,s') = 250$. The available actions from state s' are $down$ and $stay$. Assuming a discount of $\gamma = 0.5$, calculate the new estimate of the Q value for s based on this episode.

4. With this new estimate and a learning rate (α) of 0.5, update the weights for each feature.

Q3. MDPs and RL

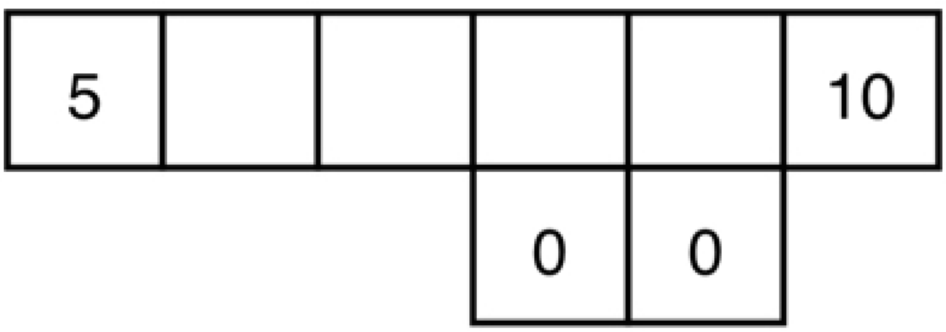


Consider the above gridworld. An agent is currently on grid cell S, and would like to collect the rewards that lie on both sides of it. If the agent is on a numbered square, its only available action is to Exit, and when it exits it gets reward equal to the number on the square. On any other (non-numbered) square, its available actions are to move East and West. Note that North and South are never available actions.

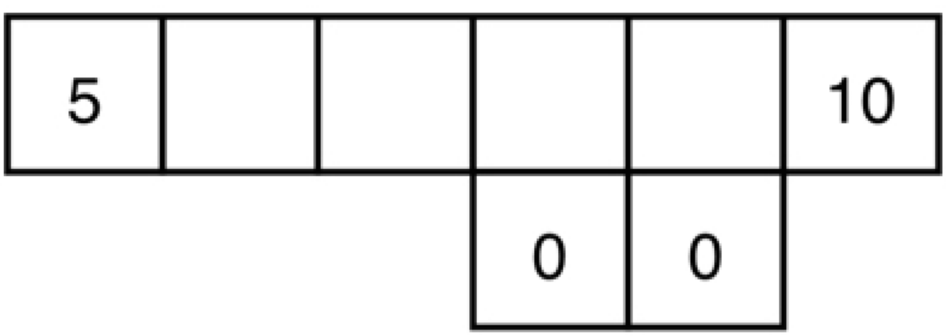
If the agent is in a square with an adjacent square downward, it does not always move successfully: when the agent is in one of these squares and takes a move action, it will only succeed with probability p. With probability $1 - p$, the move action will fail and the agent will instead move downwards. If the agent is not in a square with an adjacent space below, it will always move successfully.

For parts (a) and (b), we are using discount factor $\gamma \in [0, 1]$.

(a) Consider the policy π_{East}, which is to always move East (right) when possible, and to Exit when that is the only available action. For each non-numbered state x in the diagram below, fill in $V^{\pi_{\text{East}}}(x)$ in terms of γ and p.



(b) Consider the policy π_{West}, which is to always move West (left) when possible, and to Exit when that is the only available action. For each non-numbered state x in the diagram below, fill in $V^{\pi_{\text{West}}}(x)$ in terms of γ and p.



(c) For what range of values of p *in terms of* γ is it optimal for the agent to go West (left) from the start state (S)?

Range: ____________________

(d) For what range of values of p *in terms of* γ is π_{West} the optimal policy?

Range: ____________________

(e) For what range of values of p *in terms of* γ is π_{East} the optimal policy?

Range: ____________________