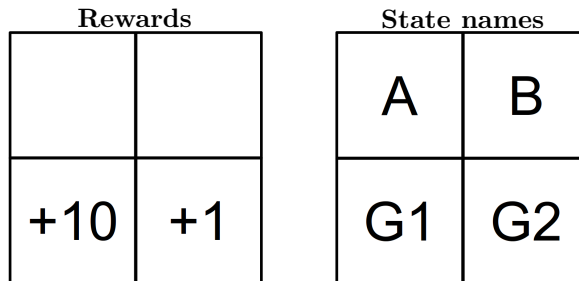


1 Q-learning

Consider the following gridworld (rewards shown on left, state names shown on right).



From state A, the possible actions are right(\rightarrow) and down(\downarrow). From state B, the possible actions are left(\leftarrow) and down(\downarrow). For a numbered state (G1, G2), the only action is to exit. Upon exiting from a numbered square we collect the reward specified by the number on the square and enter the end-of-game absorbing state X . We also know that the discount factor $\gamma = 1$, and in this MDP all actions are **deterministic** and always succeed.

Consider the following episodes:

Episode 1 ($E1$)				Episode 2 ($E2$)				Episode 3 ($E3$)				Episode 4 ($E4$)			
s	a	s'	r	s	a	s'	r	s	a	s'	r	s	a	s'	r
A	\downarrow	G1	0	B	\downarrow	G2	0	A	\rightarrow	B	0	B	\leftarrow	A	0
G1	exit	X	10	G2	exit	X	1	B	\downarrow	G2	0	A	\downarrow	G1	0
								G2	exit	X	1	G1	exit	X	10

- (a) Consider using temporal-difference learning to learn $V(s)$. When running TD-learning, all values are initialized to zero.

For which sequences of episodes, if repeated infinitely often, does $V(s)$ converge to $V^*(s)$ for all states s ?

(Assume appropriate learning rates such that all values converge.)

Write the correct sequence under "Other" if no correct sequences of episodes are listed.

- | | | | |
|---|---|---|---|
| <input type="checkbox"/> $E1, E2, E3, E4$ | <input type="checkbox"/> $E1, E2, E1, E2$ | <input type="checkbox"/> $E1, E2, E3, E1$ | <input type="checkbox"/> $E4, E4, E4, E4$ |
| <input type="checkbox"/> $E4, E3, E2, E1$ | <input type="checkbox"/> $E3, E4, E3, E4$ | <input type="checkbox"/> $E1, E2, E4, E1$ | |

Other _____

- (b) Consider using Q-learning to learn $Q(s, a)$. When running Q-learning, all values are initialized to zero. For which sequences of episodes, if repeated infinitely often, does $Q(s, a)$ converge to $Q^*(s, a)$ for all state-action pairs (s, a)

(Assume appropriate learning rates such that all Q-values converge.)

Write the correct sequence under “Other” if no correct sequences of episodes are listed.

- $E1, E2, E3, E4$ $E1, E2, E1, E2$ $E1, E2, E3, E1$ $E4, E4, E4, E4$
 $E4, E3, E2, E1$ $E3, E4, E3, E4$ $E1, E2, E4, E1$

Other _____

2 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.

- (a) 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

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

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

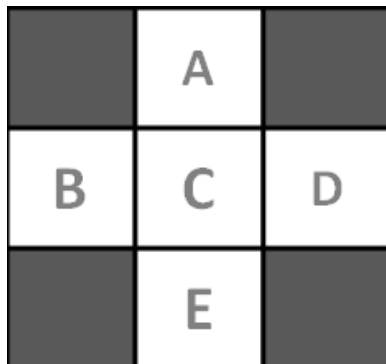


calculate the features for the actions $\in \{left, right, up, stay\}$

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

3 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 $\alpha = 0.5$.

(a) What are the learned values from TD learning after all four observations?

(b) What are the learned Q-values from Q-learning after all four observations?