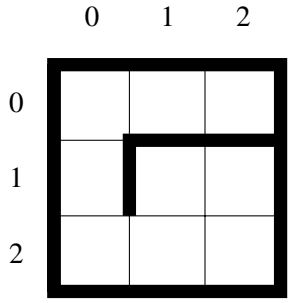
1 Friendlier Pacman

Pacman is stuck in a friendlier maze where he gets a reward every time he visits state (0,0). The precise reward function is: $R_{(0,0),a} = 1$ for any action a and $R_{s',a} = 0$ for all $s' \neq (0,0)$. This setup is a bit different from the one you've seen before: Pacman can get the reward multiple times; these rewards do not get "used up" like food pellets and there are no "living rewards". As usual, Pacman can not move through walls and may take any of the following actions: go North (↑), South (↓), East (→), West (←), or stay in place (∘). State (0,0) gives a total reward of 1 **every time** Pacman takes an action in that state regardless of the outcome, and all other states give no reward.

You should not need to use any other complicated algorithm/calculations to answer the questions below. We remind you that geometric series converge as follows: $1 + \gamma + \gamma^2 + \cdots = 1/(1-\gamma)$.

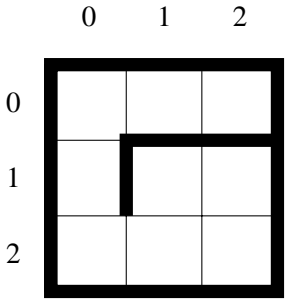
1. Assume finite horizon of 8 (so Pacman takes exactly 8 actions) and no discounting ($\gamma = 1$). Now, for each state...

...fill an the optimal policy:



(in each box draw one of the available actions: ↑, ↓, →, ←, ∘)

...and fill in the corresponding value (utility):



Hint: given each sequence of states and actions how many points can the agent get, how many times can it play any action in (0,0)?

2. Given the problem set up described above, and the utility from answer 1, what are the following Q-values?

   (a) The Q value of state-action $(0,0)$, (South) is: _____

   (b) The Q value of state-action $(2,2)$, (North) is: _____

3. Assume finite horizon of 8, no discounting, but the action to stay in place is temporarily (for this sub-point only) unavailable. Actions that would make Pacman hit a wall are not available. Specifically, Pacman can not use actions North or West to remain in state $(0,0)$ once he is there.

   (a) *True or False:* There is just one optimal action at state $(0,0)$: _____

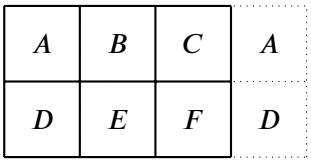
   (b) The value of state $(0,0)$ is: _____

4. Now assume infinite horizon and a discount factor $\gamma = 0.875$.

   The value of state $(0,0)$ is: _____

2 Learning by Example

Consider the following MDP with state space $S = \{A, B, C, D, E, F\}$ and action space $\mathcal{A} = \{left, right, up, down, stay\}$. Notice that C and F and connect to A and D respectively. However, we do not know the transition dynamics or reward function (we do not know what the resulting next state and reward are after applying an action in a state).



1. We are now given a policy π and would like to determine how good it is using Temporal Difference Learning with $\alpha = 0.25$ and $\gamma = 1$. We run it in the environment and observe the following transitions. After observing each transition, we update the value function, which is initially 0. Fill in the blanks with the corresponding values of the Utility function after these updates.

Episode Number	State	Action	Reward	Next State
1	A	right	12	B
2	B	right	4	C
3	B	down	−12	E
4	C	down	−16	F
5	F	stay	4	F
6	C	down	−9	F

State	$U^{\pi}(state)$
A	
B	
C	
D	
E	
F	

3 Reinforcements

Consider an unknown MDP with three states (A, B and C) and two actions ($\leftarrow$ and $\rightarrow$). Suppose the agent chooses actions according to some policy π in the unknown MDP, collecting a dataset consisting of samples (s, a, s', r) representing taking action a in state s resulting in a transition to state s' and a reward of r.

s	a	s'	r
A	$\rightarrow$	B	4
C	$\leftarrow$	B	4
B	$\rightarrow$	C	−4
A	$\rightarrow$	B	6

You may assume a discount factor of $\gamma = 1$.

1. Recall the update function of Q-learning is:

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_t + \gamma \max_{a'} Q(s_{t+1}, a')\right)$$

Assume that all Q-values are initialized to 0, and use a learning rate of $\alpha = \frac{1}{2}$.

(a) Run Q-learning on the above episode table and fill in the following Q-values:

$Q(A, \rightarrow) =$ _____ $Q(B, \rightarrow) =$ _____

(b) After running Q-learning and producing the above Q-values, you construct a policy π_Q that maximizes the Q-value in a given state:

$$\pi_Q(s) = \arg\max_a Q(s, a).$$

What are the actions chosen by the policy in states A and B?

$\pi_Q(A)$ is equal to:

- ○ $\pi_Q(A) = \leftarrow$.
- ○ $\pi_Q(A) = \rightarrow$.
- ○ $\pi_Q(A) =$ Undefined.

$\pi_Q(B)$ is equal to:

- ○ $\pi_Q(B) = \leftarrow$.
- ○ $\pi_Q(B) = \rightarrow$.
- ○ $\pi_Q(B) =$ Undefined.

2. Use the empirical frequency count model-based reinforcement learning method described in lectures to estimate the transition function $\hat{T}(s, a, s')$ and reward function $\hat{R}(s, a, s')$. (Do not use pseudocounts; if a transition is not observed, it has a count of 0.)

Write down the following quantities. You may write N/A for undefined quantities.

$\hat{T}(A, \rightarrow, B) =$ _____ $\hat{R}(A, \rightarrow, B) =$ _____

$\hat{T}(C, \leftarrow, B) =$ _____ $\hat{R}(C, \leftarrow, B) =$ _____

$\hat{T}(B, \rightarrow, A) =$ _____ $\hat{R}(B, \rightarrow, A) =$ _____

$\hat{T}(B, \leftarrow, A) =$ _____ $\hat{R}(B, \leftarrow, A) =$ _____

4 Reinforcement Learning Background

1. Each True/False question is worth 1 points. *Briefly justify* your answers.

(i) [*true* or *false*] Temporal difference learning is an offline learning method.

(ii) [*true* or *false*] Q-learning: Using an optimal exploration function can lead to a chance of regret while learning the optimal policy.

(iii) [*true* or *false*] In a deterministic MDP (i.e. one in which each state / action leads to a single deterministic next state), the Q-learning update with a learning rate of $\alpha = 1$ will correctly learn the optimal q-values (assume that all state/action pairs are visited sufficiently often).

(iv) [*true* or *false*] A large discount (close to 1) encourages greedy behavior.

(v) [*true* or *false*] A large, negative living reward ($\ll 0$) encourages greedy behavior.

(vi) [*true* or *false*] A negative living reward can always be expressed using a discount < 1.

2. This question considers properties of reinforcement learning algorithms for *arbitrary* discrete MDPs.

   (a) Select all the following methods which, at convergence, do not provide enough information to obtain an optimal policy. (Assume adequate exploration.)

   - ☐ Model-based learning of $T(s, a, s')$ and $R(s, a, s')$.
   - ☐ Direct Evaluation to estimate $U(s)$.
   - ☐ Temporal Difference learning to estimate $U(s)$.
   - ☐ Q-Learning to estimate $Q(s, a)$.

   (b) In the limit of infinite timesteps, select all of the following exploration policies for which Q-learning is guaranteed to converge to the optimal Q-values for all state. (You may assume the learning rate α is chosen appropriately, and that the MDP is ergodic: i.e., every state is reachable from every other state with non-zero probability.)

   - ☐ A fixed policy taking actions uniformly at random.
   - ☐ A greedy policy.
   - ☐ An ϵ-greedy policy
   - ☐ A fixed optimal policy.