1 RL: What's Changed Since MDPs?

1. Recall the Bellman Equation we used in MDPs to determine the value of a given state:

$$V^*(s) = \max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V(s')]$$

   What information do we no longer have direct access to in the transition to RL?

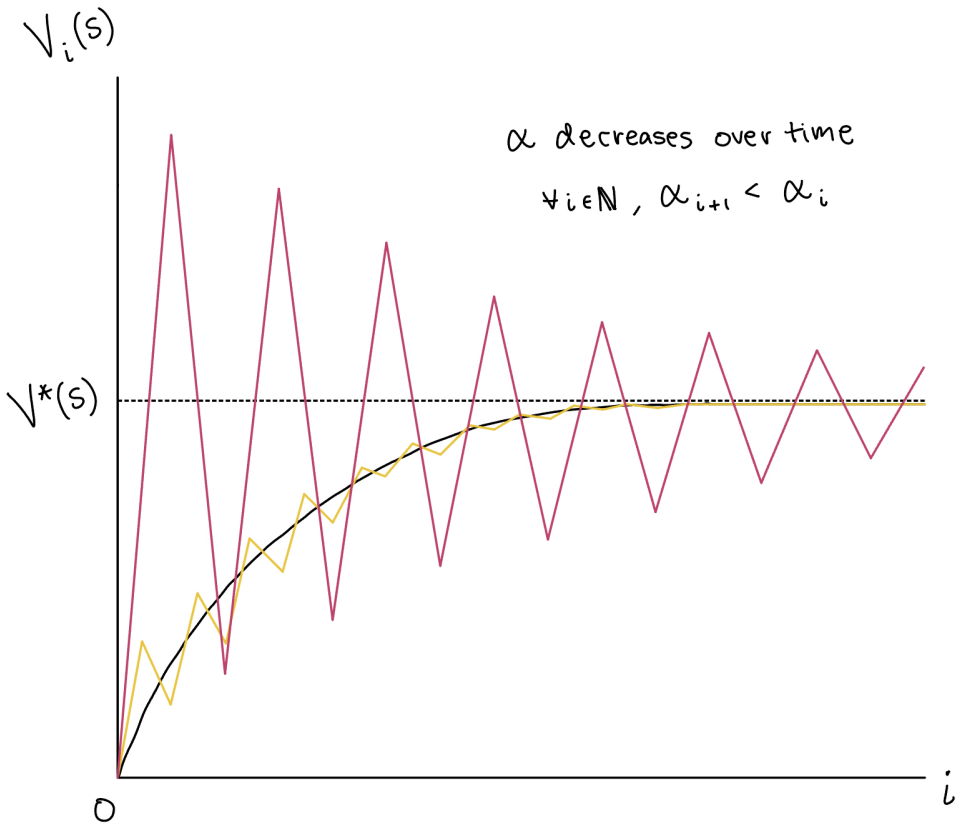
2. What is the difference between online and offline learning? Which type of learning does MDP use? How about RL?

2 RL: Conceptual Questions

Recall that in Q-learning, we continually update the values of each Q-state by learning through a series of episodes, ultimately converging upon the optimal policy.

1. What's the main shortcoming of TD learning that Q-learning resolves?

2. We are given two runs of TD-learning using the same sequence of samples but different α values depicted in the plot below. Assume the dashed horizontal line represents the optimal value for a specific state s and the black curve represents the smoothest transition to the optimal value given this sequence of samples. In both runs α decreases over time (or iterations), but one run has α values larger than the other run at any point in time. Which run (red or yellow) corresponds to the smaller values of α? How do the relative sizes of α affect the rate of convergence to the optimal value?



3. We are given a pre-existing table of current estimate of Q-values (and its corresponding policy), and asked to perform ϵ-greedy Q-learning. Individually, what effect does setting each of the following constants to 0 have on this process?
   Remember that in ϵ-greedy Q-learning, we follow the following formulation for choosing our action:

$$\text{action at time } t = \begin{cases} \arg\max_{Q(s,a)} & \text{with probability } 1-\epsilon \\ \text{any action } a & \text{with probability } \epsilon \end{cases}$$

   (a) α
   (b) γ
   (c) ϵ

4. Consider a variant of the ϵ-greedy Q-learning algorithm that is changed such that instead of using the policy extracted from our current Q-values, we use a fixed policy instead. We still perform exploration with probability ϵ. If this fixed policy happens to be optimal, how does the performance of this algorithm compare to normal ϵ-greedy Q-learning?

5. Recall the count exploration function used in the modified Q-update:

$$f(u,n) = u + \frac{k}{n+1}$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} f(Q(s',a'), N(s',a')) - Q(s,a)]$$

   Remember that k is a hyperparameter that the designer chooses, and $N(s',a')$ is the number of times we've visited the (s',a') pair. What is the effect of increasing or decreasing k?

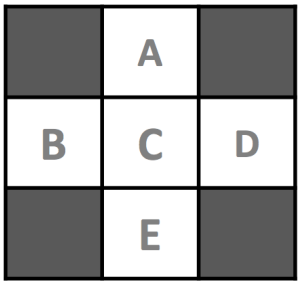
6. Let's revisit the Nim code. What RL strategies does `AgentRL` employ? Does it evaluate states or Q-states?

7. Contrast the following pairs of reinforcement learning terms:

   (i) Off-policy vs. on-policy learning
   (ii) Model-based vs. model-free
   (iii) Passive vs. active

3 Temporal Difference Learning and Q-Learning

Consider the Gridworld example that we looked at in lecture. We would like to use TD learning to find the values of these states.



Suppose we use an ϵ-greedy policy and observe the following $(s, a, s', R(s,a,s'))^*$ transitions and rewards:

$$(B, \text{East}, C, 2), (C, \text{South}, E, 4), (C, \text{East}, A, 6), (B, \text{East}, C, 2)$$

**Note that the $R(s, a, s')$ in this notation refers to observed reward, not a reward value computed from a reward function (because we don't have access to the reward function).*

The initial value of each state is 0. Let $\gamma = 1$ and $\alpha = 0.5$.

1. What are the learned values for each state from TD learning after all four observations?
2. In class, we presented the following two formulations for TD-learning:

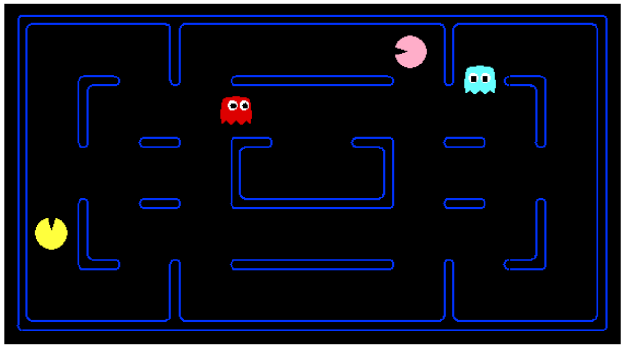
$$V^\pi(s) \leftarrow (1-\alpha)V^\pi(s) + (\alpha)sample$$

$$V^\pi(s) \leftarrow V^\pi(s) + \alpha(sample - V^\pi(s))$$

   Mathematically, these two equations are equivalent. However, they represent two conceptually different ways of understanding TD value updates. How could we intuitively explain each of these equations?
3. What are the learned Q-values from Q-learning after all four observations? Use the same $\alpha = 0.5, \gamma = 1$ as before.

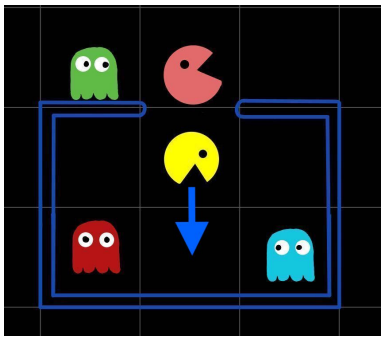
4 Approximate Q-Learning

Adabelle and Elizabeth are training agents to play AI eTag, which is totally different from Pacman. In this game, the player must find the other player in a maze. However, there are phantoms (note: these are not ghosts) that both players must avoid. However, it is unknown what the scores are for staying alive, for being caught by a phantom, or for finding the other player. Here's a sample board:



We want to apply Q-learning to this game to train our agents, but it takes a lot of memory to hold the entire grid. To remedy this, we switch to feature-based representation.

1. What features would you extract from an eTag board to judge the expected outcome of the game?
2. Say our two features are the number of phantoms in one step of our agent (F_p) and the Manhattan distance to our friend (F_d). Consider the state from the perspective of the yellow agent (the bottom agent). What are the feature values for the following board and action (note: the gridlines are drawn to help measure distance by eye):



3. When we get to this state, we have learned weights $w_p = 100$ and $w_d = 10$. Using a linear approximator, calculate the approximate Q-value for the state and action above.

4. We have received an episode, which is a start state s, the action $South$, and an end state s', and a reward $R(s, South, s')$. Now, we must update the values. The start state is the state above, and the next state has feature values $F_p = 3$ and $F_d = 1$, and the reward was 20. Assuming discount factor of 0.5, calculate the new estimate of the Q-value for s' **based on the sample**, i.e. $R(s, South, s') + \gamma \max_{a'} Q(s', a')$.

5. Now let's update the weights for each feature, given that our learning rate α is 0.5.

6. At a high-level, what did our approximate Q-learning agent learn here? Once we finish learning, how can we evaluate our agents' performances?