1 RL: What's Changed Since MDPs?

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

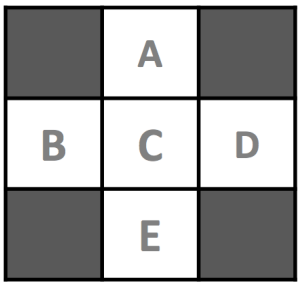
$$V^*(s) = \max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V(s')]$$

   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 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?

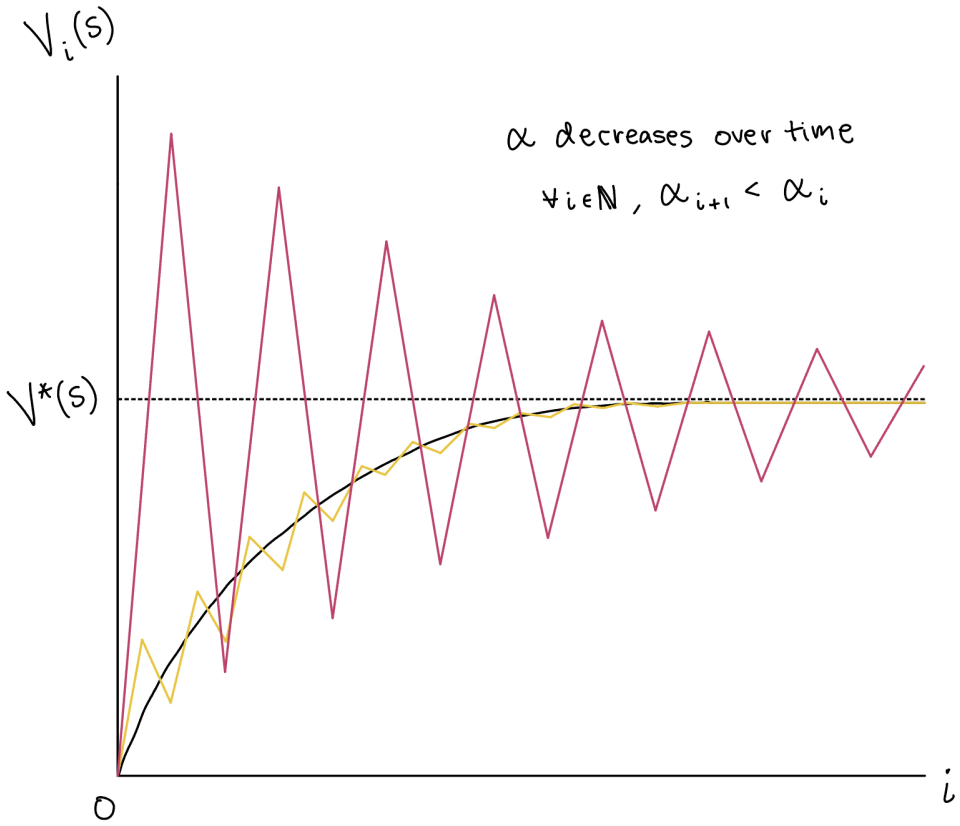
3. What are the learned Q-values from Q-learning after all four observations? Use the same $\alpha = 0.5, \gamma = 1$ as before.

3 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:

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

   (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. 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

4 Approximate Q-Learning

You decide to go to Kennywood this weekend. Our RL problem is based on choosing a ride from a set of many rides.

You start off feeling well, getting positive rewards from rides, some larger than others. However, there is some chance of each ride making you sick. If you continue going on rides while sick there is some chance of becoming well again, but you don't enjoy the rides as much, receiving lower rewards (possibly negative).

You have never been to an amusement park before, so you don't know how much reward you will get from each ride (while well or sick). You also don't know how likely you are to get sick on each ride, or how likely you are to become well again. What you do know about the rides is:

Actions / Rides	**Type**	**Wait**	**Speed**
Steel Curtain	Rollercoaster	Long	Fast
Lil' Phantom	Rollercoaster	Short	Slow
Cranky's Tower	Drop Tower	Short	Fast
Pirate	Pendulum	Short	Slow
Leave the Park	Leave	Short	Slow

We will formulate this as an RL problem with two states, **well** and **sick**. Each ride corresponds to an action. The 'Leave the Park' action ends the current run through the problem. Taking a ride will lead back to the same state with some probability or take you to the other state. We will use a feature-based approximation to the Q-values, defined by the following four features and associated weights:

Features	**Initial Weights**
$f_0 = f_{\text{sick}}(s,a) = \begin{cases} 0, s = \text{Well} \\ -5, s = \text{Sick} \end{cases}$	$w_0 = 1$
$f_1 = f_{\text{type}}(s,a) = \begin{cases} 1, \text{ if } \textit{action} \text{ type is Rollercoaster} \\ 0, \text{otherwise} \end{cases}$	$w_1 = 2$
$f_2 = f_{\text{wait}}(s,a) = \begin{cases} 1, \text{ if } \textit{action} \text{ wait is Short} \\ 0, \text{otherwise} \end{cases}$	$w_2 = 1$
$f_3 = f_{\text{speed}}(s,a) = \begin{cases} 1, \text{ if } \textit{action} \text{ speed is Fast} \\ 0, \text{otherwise} \end{cases}$	$w_3 = 0.5$

1. Calculate Q-values for each action, given the state is $'Sick'$.
2. Apply a Q-learning update based on the sample ('Well', 'Steel Curtain', 'Sick', -10.5), using a learning rate of $\alpha = 0.5$ and discount of $\gamma = 0.5$. What are the new weights?
3. Now we will consider the exploration exploitation tradeoff in this amusement park. Assume we have the initial weights from the table above. What action will an ϵ-greedy approach choose from the well state? If multiple actions could be chosen, give each action and its probability.
4. When running Q-learning another approach to dealing with this tradeoff is using an exploration function:

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

   (a) How is this function used in the Q-learning equations?

   (b) What does u represent in the exploration function?

   (c) What does n represent in the exploration function?

   (d) What does k represent in the exploration function?