CS 188 Spring 2024 Introduction to Artificial Intelligence

Exam Prep 12 Solutions

Q1. CalDining Bandits

You're an excited new student who wants to know where to eat lunch at Berkeley! Every day at lunchtime, you take action a to use your meal swipe at Crossroads ($a = X$), Cafe 3 ($a = C$), or Golden Bear Cafe ($a = G$) (the other dining halls are too inconvenient). Let a_i be the action you take on day i.

Suppose that the reward you get from croads (X) is uniformly distributed between -10 and 50, the reward you get from Cafe 3 (C) is uniformly distributed between 0 and 30, and the reward you get from GBC (G) is always 15.

(a) What is the optimal value V^*? Which dining hall has the best expected reward?

$V^* = \text{argmax}_a E(r|a) = \boxed{20}$

The best action is to go to croads (hot take).

(b) What is the gap Δ_C for the action of going to Cafe 3 (C)?

$Q(C) = E(r|C) = 15$

$\Delta_C = V^* - Q(C) = 5$

(c) Suppose Cafe 3 just happens to be right next to your dorm, so your policy is to always choose action C. What is the **regret** l_t for one action under this policy?

$l_t = E[V^* - Q(a_t)] = V^* - Q(C) = 5$

(d) Now suppose you are indecisive, so your policy is to randomly choose a dining hall to go to each day. What is the **regret** l_t for one action under this policy?

$l_t = E[V^* - Q(a_t)]$

$= \frac{1}{3}(V^* - Q(X)) + \frac{1}{3}(V^* - Q(C)) + \frac{1}{3}(V^* - Q(G))$

$= 0 + \frac{5}{3} + \frac{2}{3}$

$= \boxed{\frac{7}{3}}$

(e) Suppose you follow the random policy from the previous part for 5 days, taking actions X, C, C, G, X and getting rewards $10, 20, 22, 18, -10$. What is the **total regret** for this policy? (Hint: Trick question?)

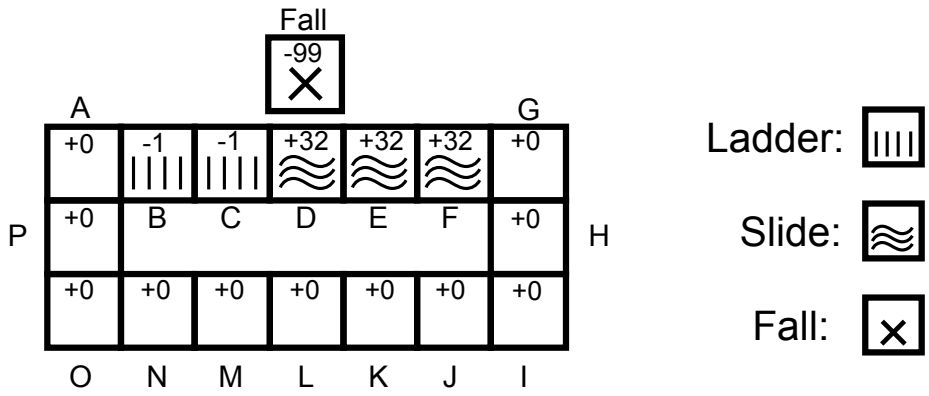
In this class, regret is used to refer to "expected suboptimality", and total regret is also an expectation. As such, the total regret is 5 times the result from the previous part, so

$L_5 = \boxed{\frac{35}{3}}$

(f) True or False: Using the UCB1 algorithm for this problem would lead to logarithmic total regret, after enough days.

True, taken directly from lecture slides.

Q2. RL: Dangerous Water Slide



Suppose now that several years have passed and the water park has not received adequate maintenance. It has become a dangerous water park! Now, each time you choose to move to (or remain at) one of the ladder or slide states (states B-F) there is a chance that, instead of ending up where you intended, you fall off the slide and hurt yourself. The cost of falling is -99 and results in you getting removed from the water park via ambulance. The new MDP is depicted above.

Unfortunately, you don't know how likely you are to fall if you choose to use the slide, and therefore you're not sure whether the fun of the ride outweighs the potential harm. You use reinforcement learning to figure it out!

For the rest of this problem assume $\gamma = 1.0$ (i.e. no future reward discounting). You will use the following two trajectories through the state space to perform your updates. Each trajectory is a sequence of samples, each with the following form: (s, a, s', r).

Trajectory 1: (A, East, B, -1), (B, East, C, -1), (C, East, D, +32)
Trajectory 2: (A, East, B, -1), (B, East, Fall, -99)

(a) What are the values of states A, B, and C after performing temporal difference learning with a learning rate of $\alpha = 0.5$ using only Trajectory 1?

$V(A) =$ -0.5 $V(B) =$ -0.5 $V(C) =$ 16

(b) What are the values of states A, B, and C after performing temporal difference learning with a learning rate of $\alpha = 0.5$ using both Trajectory 1 and Trajectory 2?

$V(A) =$ -1.0 $V(B) =$ -49.75 $V(C) =$ 16

(c) What are the values of states/action pairs (A, South), (A, East), and (B, East) after performing Q-learning with a learning rate of $\alpha = 0.5$ using both Trajectory 1 and Trajectory 2?

$Q(A, \text{South}) =$ 0.0 $Q(A, \text{East}) =$ -0.75 $Q(B, \text{East}) =$ -49.75