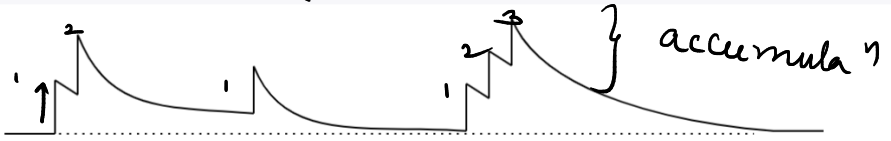
RL - Week 9

GA

Consider the following diagram that displays the eligibility trace for a state s in some episode:



1) What kind of eligibility trace is used here? *1 point*

- Accumulating trace
- Replacing trace

2) How many times is state s visited in this episode? = total no. of peaks = 2+1+3 = 6

You are training an agent with TD(λ) algorithm. There are a total of 10 states s_i for $i \in [0, 8]$ and a terminal state s_T. Following is the first trajectory the agent observes:

$state = s_0 \rightarrow action = a_0 \rightarrow reward = 0 \rightarrow$

$state = s_1 \rightarrow action = a_1 \rightarrow reward = 0 \rightarrow$

$state = s_1 \rightarrow action = a_2 \rightarrow reward = 0 \rightarrow$

$state = s_2 \rightarrow action = a_3 \rightarrow reward = +10 \rightarrow state = s_T$

$s_0, a_0 \xrightarrow{0} s_1, a_1 \xrightarrow{0} s_1, a_2 \xrightarrow{0} s_2, a_3 \xrightarrow{10} s_T$

$\gamma = 1$

$\lambda = 0.9$

$\alpha = 1$

Assume the following:

discount factor ($\gamma = 1$)
Lambda ($\lambda = 0.9$)
Learning rate ($\alpha = 1$)
$V(s)$ is initialized to $0 \; \forall \; s$
s_T is a terminal state
The eligibility trace of a state is denoted by $e(s)$.

$V_s = 0$

$s_T = 0$

$e(s)$

3) What will be the eligibility trace of state s_4 once the episode concludes but before the next episode begins? = 0

s_0, s_1, s_1, s_2, s_T

since s_4 is not a part of this trajectory, the eligi trace of s_4, i.e $e(s_4)$ is affected by decay factor. Initially, all e_s are set to 0. $\therefore$, $e(s_4) = 0$.

4) What will be the eligibility trace of state s_0 once the episode concludes but before the next episode begins? *1 point*

0.729

s_0 is visited at first time step. It doesn't get repeated

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) 0.729

$s_0 \xrightarrow[1]{} s_1 \xrightarrow[0.9]{} s_1 \xrightarrow[0.9]{} s_2 \xrightarrow[0.9]{} s_f$

$(0.9)^3 = 0.729$

5) What will be the eligibility trace of state s_1 – if eligibility trace is accumulating – once the episode concludes but before the next episode begins? *1 point*

1.71

$s_0 \Rightarrow 0$

$s_1 \Rightarrow 0.9 \times 0 = 0$

$s_1 \Rightarrow 0.9 \times 0 + 1 = 1$

$s_2 \Rightarrow 0.9 \times 1 + 1 = 1.9$

$s_f \Rightarrow 0.9 \times 1.9 = 1.71$

6) What will be the eligibility trace of state s_1 — if eligibility trace is replaced for a state encountered in an episode — once the episode concludes but before the next episode begins?

$S_0 = 0$

$S_1 = 0.9 \times 0 = 0$

$S_1 = 1$

$S_2 = 1$

$S_t = 0.9 \times 1 = 0.9$

0.9

Yes, the answer is correct.

Score: 1

Accepted Answers:

(Type: Numeric) 0.9

1 point

7) Consider a binary bandit, with policy described as follows:

$$\pi(a,\theta) = \begin{cases} \theta, & \text{if a=1} \\ 1-\theta, & \text{if a=0} \end{cases}$$

At the beginning $\theta = 0.5$. What will be probability of pulling arm $a = 1$ after pulling arm $a = 0$ and receiving reward of $+2$? Assume baseline to be 0 and learning rate (ρ) to be 0.01.

0.49

$\theta_0 = 0.5$

$\pi(a,\theta) = \begin{cases} \theta & a=1 \\ 1-\theta & a=0 \end{cases}$

$(a=1 \mid a=0), \; r=2$

$\rho = 0.01$

$b = 0$

$\Delta\theta = \rho (r_t - b_t)(a-\theta)$

$= 0.01 (2-0)(0-0.5) = -0.01$

$\theta' = \theta + \Delta\theta$

$= 0.5 + (-0.01) = 0.49$

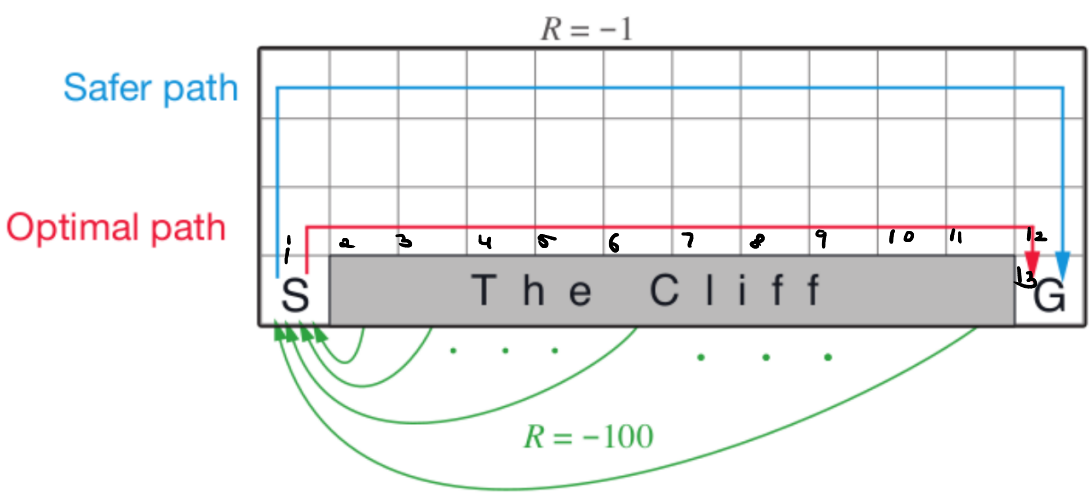
8) If action space is continuous, and taking action a is represented by a normal distribution with parameters μ and σ, that is, $a \sim \mathcal{N}(\mu, \sigma^2)$, which of the following **1 point** is the correct update rule for updating parameters μ and σ? Use a common learning rate for updating both parameters. Specifically, let the learning rates be $\alpha_\mu = \alpha_\sigma = \alpha\sigma^2$.

$\ln \pi = C - \ln \sigma - \frac{(a-\mu)^2}{2\sigma^2}$

$\frac{\delta \ln \pi}{\delta \ln \sigma} = \frac{a-\mu}{\sigma^2}$

- $\Delta\mu = \frac{\alpha}{\sigma^2}(r-\bar{r})(a-\mu)$
  $\Delta\sigma = \frac{\alpha}{\sigma^3}(r-\bar{r})\left[(a-\mu)^2 - \sigma^2\right]$
- $\Delta\mu = \alpha(r-\bar{r})(a-\mu)$
  $\Delta\sigma = \alpha(r-\bar{r})\left[(a-\mu)^2 - \sigma^2\right]$
- $\Delta\mu = \frac{\alpha}{\sigma^3}(r-\bar{r})(a-\mu)$
  $\Delta\sigma = \frac{\alpha}{\sigma^3}(r-\bar{r})\left[(a-\mu)^2 - \sigma^2\right]$
- $\Delta\mu = \alpha(r-\bar{r})(a-\mu)$
  $\Delta\sigma = \frac{\alpha}{\sigma}(r-\bar{r})\left[(a-\mu)^2 - \sigma^2\right]$

Consider the cliff walking task. All transitions are deterministic. The reward is -1 on all transitions except those that take the agent into the cliff region. Any action that takes the agent into the cliff region results in a reward of -100 and the agent is transported back to the start state. $\gamma = 1$ for this task.

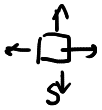


The action values for SARSA and Q-learning are learnt over $10,000$ episodes. Assume that the Q values converge at the end of these many episodes. For both algorithms, $\epsilon = 0.1$ and is not changed throughout the learning, $\alpha = 0.5$ and is steadily decreased over time.

9) In the case of SARSA, the Q values for the state just above the start state is given below: 1 point

$-20, \quad -26, \quad -26, \quad -33$

From left to right, what could be the actions corresponding to these Q values?



SARSA will always learn the safest path. which is up. (-20)

- LEFT, RIGHT, UP, DOWN
- ✓ UP, DOWN, LEFT, RIGHT
- RIGHT, UP, LEFT, DOWN
- DOWN, LEFT, RIGHT, UP

-33 is right because it takes you closer to cliff, dangerous path.

10) Find the value of $Q(S, \text{up})$ in the case of Q-learning, where S is the start state. Enter the nearest integer as your answer.

Hint: Q-learning learns the optimal policy for this task.

-13

Q-learning will choose the most optimal path close to the cliff. So, it takes 13 steps to reach the goal.

(13 x -1)

↙ steps ↘ reward