

RL1: Bandits

Q.1 K-armed bandit $\rightarrow k=4$.

Using ϵ -greedy action selection algorithm,

Given: $A_1 = 1, R_1 = -1$ (prob -3 of getting -1)

$A_2 = 2, R_2 = 0$ (prob 2 of getting 0)

$A_3 = 2, R_3 = -2$

$A_4 = 2, R_4 = 2$

$A_5 = 3, R_5 = 0$

Q: On which timestep ϵ -case occurred \rightarrow 1) definitely?
2) May be?

$$\rightarrow \textcircled{1} Q_t(a) = \frac{\sum (\text{Rewards when a taken prior to } t)}{\text{No. of times a taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

$$\textcircled{1} A_0 = 0, R_0 = 0 \quad \boxed{Q_1 = 0}, Q_2 = 0, Q_3 = 0, Q_4 = 0 \\ \text{choose 1 (Maybe Random)} \quad \text{As } Q_1 = Q_2 = Q_3 = Q_4 = 0.$$

$$\textcircled{2} A_1 = 1, R_1 = -1 \quad Q_1 = \frac{-1}{1} = -1 \quad \boxed{Q_2 = 0}, Q_3 = 0, Q_4 = 0.$$

Now for $t=2$, We got Reward = -ve. from A_1 (action 1) we continue to exploit initial expected val for '0' $\rightarrow A_2$

Choose 2, May be Random

$$Q_1 = -1 < Q_2 = Q_3 = Q_4 = 0.$$

\rightarrow consistent with ϵ -greedy selection.

$$\textcircled{3} A_2 = 2, R_2 = +1 \quad Q_1 = -1, Q_2 = \frac{1}{1} = 1 \quad Q_3 = 0, Q_4 = 0.$$

\rightarrow Action A_2 selected to exploit ϵ

\rightarrow Maybe random selection Action 2 again (^{agent} decides to exploit again)
As $\rightarrow Q_2 = 1 > Q_3 = Q_4 = 0 > Q_1 = -1$

$$④ A_3 = 2, R = -2, \varphi_1 = -1, \varphi_2 = \frac{1 + (-2)}{2} = -0.5$$

$$\varphi_3 = 0, \varphi_4 = 0.$$

$$\underline{\varphi_3 = \varphi_4 = 0 > \varphi_2 = -0.5 > \varphi_1 = -1}$$

Now, according to ϵ -greedy action selection → it should select A_3 OR A_4 .

Here, even though it should select A_3/A_4 , agent decides Action 2 again. This step is → definitely exploratory.

$$⑤ A_1 = 2, R_4 = 2, \varphi_1 = -1, \varphi_2 = \frac{1 + 2 + 2}{3} = \frac{1}{3} = 0.33$$

[Agent could have exploited action A_1/A_2]

According to ϵ -greedy action selection Algo → it should select action 2 again.

But agent definitely decides to explore A_3 (Random).

$$⑥ A_3 = 2, R_5 = 0, \varphi_1 = -1, \varphi_2 = 0.33, \varphi_3 = \frac{0}{1} = 0$$

$$\varphi_4 = 0$$

Thus, agent definitely explored A_4 & A_5 at $t=4$ & $t=5$.

But, it could have explored on (A_1, A_2, A_3) at $t=1, 2, 3$

timestamps as well.

$$0 = \varphi_1, 0 = \varphi_2, 1 = \frac{1}{1} = \varphi_3, 1 = 1, 1 + 0.33, 1 = 2A$$

→ if always at batches \leftarrow A \leftarrow

→ if always \leftarrow A \leftarrow

$$1 = 1, 1 < 0 = \varphi_1 = \varphi_2 < 1 = 2A \leftarrow A$$

Q.2

α_n = step-size parameter = non-constant

→ When $\alpha = \text{const.}$

q -value estimate for reward is given by —

$$q_{n+1} = (1-\alpha) q_n + \sum_{i=1}^n \alpha (1-\alpha)^{n-i} r_i$$

Now that step-size isn't const.

Let α_n = stepsize at each step n. from eq 2.6.

$$q_{n+1} = q_n + \alpha [R_n - q_n]$$

$$\text{let } q_{n+1} = q_n + \alpha_n [R_n - q_n] = q_n + \alpha_n R_n - \alpha_n q_n$$

$$= \alpha_n R_n + q_n (1 - \alpha_n)$$

$$= \alpha_n R_n + (1 - \alpha_n) [q_{n-1} + \alpha_{n-1} (R_{n-1} - q_{n-1})]$$

$$= \alpha_n R_n + (1 - \alpha_n) [q_{n-1} + (\alpha_{n-1} \cdot R_{n-1}) - (\alpha_{n-1} \cdot q_{n-1})]$$

$$= \alpha_n R_n + (1 - \alpha_n) [\alpha_{n-1} \cdot R_{n-1} + q_{n-1} (1 - \alpha_{n-1})]$$

$$\text{thus } q_{n+1} = \alpha_n R_n + (1 - \alpha_n) \cdot \alpha_{n-1} \cdot R_{n-1} + (1 - \alpha_n) q_{n-1} (1 - \alpha_{n-1})$$

$$= \sum_{i=1}^n [\alpha_i R_i \cdot \prod_{j=i}^{n-1} (1 - \alpha_j)] + q_1 \left(\prod_{i=1}^n (1 - \alpha_i) \right)$$

Hence, weighting on each of the prior reward is →

$$\sum_{i=1}^n \alpha_i \prod_{j=1}^{n-1} (1 - \alpha_j) \quad \text{for every reward } R_i \in [1, n]$$

Q.3 Bias in q -value estimates: = following slide = no Q.P

a) Sample avg. estimate in eqn 2.1 \rightarrow

We say \rightarrow estimate is biased if the expected value of the estimate doesn't match the true value.
ie $E[q_n] \neq q^*$ (otherwise it's unbiased)

$$q_{t(a)} = \frac{\text{sum of rewards taken prior to 't'}}{\text{No. of times 'a' taken}} = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

$$q_n = \frac{R_1 + R_2 + R_3 + \dots + R_{n-1}}{n-1}$$

$$E[q_n] = E\left[\frac{R_1 + R_2 + R_3 + \dots + R_{n-1}}{n-1}\right] = \frac{1}{n-1} E[R_1 + R_2 + \dots + R_{n-1}]$$

$$\left[\frac{1}{n-1} (E[R_1] + E[R_2] + \dots + E[R_{n-1}]) \right]$$

$$\left[\frac{1}{n-1} (n-1) q^* \right] = q^* \rightarrow \text{Thus, it is unbiased.}$$

random var \rightarrow 1 predicate = $\begin{cases} 1 & \rightarrow \text{predicate is true} \\ 0 & \rightarrow \text{otherwise} \end{cases}$

① Hence, when will not select an action 'a' prior to 't'

b) denominator = 0 in eqn 2.1. \rightarrow $q_{t(a)}$ \rightarrow some default value.

② upon multiple trials, trial = finite large no. of action selection step.

sample avg. $q_{t(a)}$ converges to $q^*(a)$ in limit.

\leftarrow as number of trials go does not grow by law of large numbers.

b) if $q_1 = 0$ for $n > 1$: prove that $0 < \alpha < 1$ \Rightarrow 2.5.

$q_{n+1} = q_n + \alpha (R_n - q_n)$
exponential-recency weighted avg. estimate

→ from eq 2.6,

$$\varphi_{n+1} = (1-\alpha)^n \varphi_1 + \sum_{i=1}^n \alpha (1-\alpha)^{n-i} R_i$$

if $\varphi_1 = 0$ is φ_n for $n > 1$: ?

i) $\varphi_1 = 0$ & $n = 1$

$$\varphi_2 = 0 + \alpha (1-\alpha)^0 R_1$$

$$\varphi_2 = \alpha R_1 + [1 - \alpha] E[R_1] + [1 - \alpha]^2 (1 - \alpha) E[R_2] = [1 + \alpha \varphi] E[R_1]$$

e) $\varphi_1 = 0$ & $n = 2$

$$\varphi_3 = \alpha ((1-\alpha) R_1 + \alpha R_2) + \alpha (1-\alpha) E[R_3]$$

g) $n = 3$

$$\varphi_4 = \alpha ((1-\alpha)^2 R_1 + \alpha (1-\alpha) R_2 + \alpha R_3) + \alpha (1-\alpha) E[R_4]$$

$$\therefore \varphi_{n+1} = \sum_{i=1}^n \alpha (1-\alpha)^{n-i} E[R_i]$$

As $n \rightarrow \infty \rightarrow$ By large law of numbers,

$$E[R_i] = q^*$$

→ Let's solve $\alpha \rightarrow \sum_{i=1}^n \alpha (1-\alpha)^{n-i}$

$$= \alpha [1 + (1-\alpha) + (1-\alpha)^2 + \dots + (1-\alpha)^{n-1}]$$

$$= \alpha \left[1 + \frac{1-(1-\alpha)^n}{1-(1-\alpha)} \right] = \alpha \left[\frac{1-(1-\alpha)^n}{\alpha} \right]$$

$$= 1 - (1-\alpha)^n \neq 1 \quad \text{as } 0 < \alpha < 1$$

$$\therefore \varphi_{n+1} \neq q^*$$

Hence, even if $\varphi_1 = 0$, our estimate for φ_{n+1} can never reach its true value.

→ Biased.

c) derive condition for - $\hat{\theta}_n$ will be unbiased.

Ce^{-1}

$$g_{n+1} = (1-\alpha)^n g_1 + \sum_{j=1}^n \alpha(1-\alpha)^{n-j} R_i \quad \dots \text{prev. eq } n$$

for unbiased g_{n+1} ,

$$\text{we need } E[\varphi_{n+1}] = q^{*(\alpha-1)} \cdot \alpha + \alpha = \varphi$$

$$E[\varrho_{n+1}] = E[(1-\alpha)^n \varrho_1] + E\left[\sum_{j=1}^n \alpha(1-\alpha)^{n-j} r_i\right]$$

$$= (1-\alpha)^n g_1 + \sum_{j=1}^n \alpha ((1-\alpha)^{n-j}) \mathbb{E}[R_j]$$

$$= (1-\alpha)^n q_1 + \sum_{j=1}^n \alpha (1-\alpha)^{n-1-j} q^* = q^* \quad (a)$$

1.4. condition \rightarrow it is unbiased.

$$\sum_{j=1}^n \alpha(1-\alpha)^{n-j} = 1.$$

d)

d) show that Q_{n+1} is asymptotically unbiased.

$$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^n \alpha (1-\alpha)^{n-i} R_i$$

$$E[Q_{n+1}] = (1-\alpha)^n Q_1 + \sum_{i=1}^n \alpha (1-\alpha)^{n-i} E[R_i]$$

as $n \rightarrow \infty$; $0 < \alpha < 1$ $(1-\alpha)^n \rightarrow 0$

$$\rightarrow \sum_{i=1}^n \alpha (1-\alpha)^{n-i} = \alpha (1+(1-\alpha)+\dots+(1-\alpha)^{n-1})$$

$$\frac{1}{\alpha} = \alpha \left(\frac{1-(1-\alpha)^n}{1-(1-\alpha)} \right) \quad (\text{infinite sum of G.P.})$$

$$E[Q_{n+1}] = \alpha \sum_{i=1}^n 1 \cdot E[R_i] = q *$$

\rightarrow it is unbiased when $n \rightarrow \infty$

- c)
 - 1) Since Q_{n+1} depends on initial action value estimate $Q_1(a)$, \rightarrow it is generally biased.
 - 2) We might obtain large values that still lead to possibility of bias.
 - 3) exponential-averaging weighted avg. suffers from initial estimate problem.
 - 4) Unlike sample-avg. estimate, bias won't disappear, due to const value of α , it is always there, even though decreasing over time.

Q.5 Predicting Asymptotic behaviour in fig 2.2

(A) $\epsilon = 0.01 \rightarrow$ perform better → case: stationary testbed in long run

→ Reason → once it has found optimal action
it will exploit it with 0.99 chance.

→ learning curve can take more time to reach
optimal action, as the exploration prob. low.
→ But, cumulative reward = Maximum.

$$\text{Optimal Action} = \left[(-0.01) \times 1 + 0.01 \times \frac{1}{10} \right] \times 100\% = 99.1\%$$

(B) $\epsilon = 0.1 \rightarrow$ 10% explore more → it will have better
initial estimates. But it will keep on
exploiting 90% of time. so in the long run,
cumulative reward < cumulative reward
($\epsilon=0.1$) ($\epsilon=0.01$)

$$\text{optimal action} = \left[(0.1 \times 1 + 0.1 \frac{1}{10}) \right] \times 100\% = 91\%$$

(C) $\epsilon = 1$ → it will select first the result & refuse to make
new attempt.

Learning more difficult. prob bootstrap process - learning loop
- without stem cells

→ impossible to avoid estimates. prob-alignments still
correctly aligned if it follows f(x) of old
and new policies of agent new