Northeastern University
College of Engineering

EECE 5698 - ST: Reinforcement Learning Spring 2023

HW1 Solution

Problem 1.

Consider a random variable X whose pdf is:

$$P_X(x) = \begin{cases} 1/2 & x = -1 \\ 1/4 & x = 0 \\ 1/4 & x = 1 \end{cases}$$

a) Find E[X] and E[X^2]

b) Find Var[X] and σ

Solution:

a)

$S_x = \{0, -1, 1\}$

$E[X] = \sum_{x \in S_x} x \, P_X(x) = -1 \times (1/2) + 0 \times (\frac{1}{4}) + 1 \times (\frac{1}{4}) = -\frac{1}{4}$

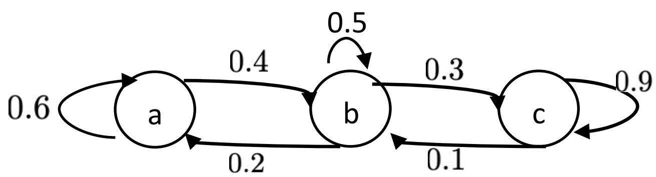
$E[x^2] = \sum_{x \in S_x} x^2 P_X(x) = (-1)^2 \times 1/2 + 0^2 \times \frac{1}{4} + (1)^2 \times \frac{1}{4} = \frac{3}{4}$

b) $Var[x] = E[x^2] - (E[x])^2 = \frac{3}{4} - (-\frac{1}{4})^2 = \frac{3}{4} - \frac{1}{16} = \frac{11}{16}$

$\sigma_x = \sqrt{Var[x]} = \frac{\sqrt{11}}{4}$

Problem 2.

Consider a Markov chain $\{x_n, n = 0, 1, \dots\}$ with a transition diagram:



a) Compute the transition matrix, given x={a,b,c}

b) Compute $p(x_k = b | x_{k-1} = a)$ and $p(x_k = b | x_{k-2} = a)$

Solution:

a)

$$M = \begin{array}{c} \\ a_{k-1}\to \\ b_{k-1}\to \\ c_{k-1}\to \end{array} \begin{array}{c} \begin{array}{ccc} a_k & b_k & c_k \end{array} \\ \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.2 & 0.5 & 0.3 \\ 0 & 0.1 & 0.9 \end{bmatrix} \end{array}$$

b)

$$P(x_k = b | x_{k-1} = a) = 0.4$$

$$P(x_k = b | x_{k-2} = a) = \sum_{i \in \{a,b\}} P(x_k = b, x_{k-1} = i | x_{k-2} = a)$$

$$= P(x_k = b, x_{k-1} = a | x_{k-2} = a) + P(x_k = b, x_{k-1} = b | x_{k-2} = b)$$

$$= P(x_k = b | x_{k-1} = a, \cancel{x_{k-2} = a}) \times P(x_{k-1} = a | x_{k-2} = a)$$

$$+ P(x_k = b | x_{k-1} = b, \cancel{x_{k-2} = b}) \times P(x_{k-1} = b | x_{k-2} = a)$$

$$= 0.4 \times 0.6 + 0.5 \times 0.4 = 0.44$$

Problem 3.

Consider two-bandit problem with the following reward distributions:

$$R(a^1) \sim Uniform[0\ \ 1.4]$$

$$R(a^2) \sim \mathcal{N}(\mu = 0.5, \sigma = 1)$$

a) Compute the optimal Q*(a¹), Q*(a²) and π^*.

b) Consider the reward distributions are unknown. Use the learning rate $\alpha = 0.5$ to estimate Q(a¹), Q(a²) and π given the following:

	k=1	k=2	k=3	k=4	k=5
Action	a^1	a^2	a^1	a^2	a^1
Reward	1	0.5	0	1.25	1.35

c) Repeat part b for optimistic initial value Given $Q(a^1) = Q(a^2) = 5$.

Solution:

Part a)

$R(a^1) \sim Uniform[0\ \ 1.4] \rightarrow E[R|a=a^1] = \frac{a+b}{2} = \frac{0+1.4}{2} = 0.7$

$R(a^2) \sim N(\mu=0.5, \sigma^2=1) \rightarrow E[R|a=a^2] = 0.5$

$Q^*(a^1) = 0.7$ and $Q^*(a^2) = 0.5$ are the optimal Q-values.

$$\pi^* = \underset{a \in \{a^1, a^2\}}{\operatorname{argmax}} Q^*(a) = a^1$$

Part b)

For $Q(a^1) = Q(a^2) = 0$, we have:

$Q(a) : Q(a) + \alpha[r - Q(a)]$

$k=1 \rightarrow Q(a^1) = 0 + 0.5[1 - 0] = 0.5$

$k=2 \rightarrow Q(a^2) = 0 + 0.5[0.5 - 0] = 0.25$

$k=3 \rightarrow Q(a^1) = 0.5 + 0.5[0 - 0.5] = 0.25$

$k=4 \rightarrow Q(a^2) = 0.25 + 0.5[1.25 - 0.25] = 0.75$

$k=5 \rightarrow Q(a^1) = 0.25 + 0.5[1.35 - 0.25] = 0.8$

$$\pi = \begin{cases} \text{Random } a^1 \text{ or } a^2 & \text{w.p. } \varepsilon \\ \underset{a \in \{a^1, a^2\}}{\operatorname{argmax}} Q(a) = a^1 & \text{w.p. } (1-\varepsilon) \end{cases}$$

Part c)

For the case with optimistic initial values $Q(a^1) = Q(a^2) = 5$, we have:

$Q(a^1) = Q(a^2) = 5$

$k=1 \rightarrow Q(a^1) = 5 + 0.5[1 - 5] = 3$

$k=2 \rightarrow Q(a^2) = 5 + 0.5[0.5 - 5] = 2.75$

$k=3 \quad Q(a^1) = 3 + 0.5[0 - 3] = 1.5$

$k=4 \rightarrow Q(a^2) = 2.75 + 0.5[1.25 - 2.75] = 2$

$k=5 \quad Q(a^1) = 1.5 + 0.5[1.35 - 1.5] = 1.425$

$$\pi = \begin{cases} \text{Random} & \text{w.p. } \varepsilon \\ \underset{a \in \{a^1, a^2\}}{\operatorname{argmax}} Q(a) = a^2 & \text{w.p. } 1-\varepsilon \end{cases}$$

Problem 4.

Given the following interaction and reward sequence, set $\alpha = 0.5$, $H_1(a^1) = H_1(a^2) = 0$ and use the gradient-bandit policy to compute $H_4(a^1)$, $H_4(a^2)$, $\pi_4(a^1)$ and $\pi_4(a^2)$.

	k=1	k=2	k=3
Action	a^1	a^2	a^1
Reward	1	0.5	0

Solution:

$H_1(a^1) = H_1(a^2) = 0$

(1) $\pi_1(a^1) = \frac{e^0}{e^0+e^0} = \frac{1}{2}$

$\pi_1(a^2) = \frac{e^0}{e^0+e^0} = \frac{1}{2}$

$\rightarrow a_1 = a^1,\ R_1 = 1 \Rightarrow \bar{R}_1 = 1$

$H_2(a^1) = H_1(a^1) + \alpha\,[R_1 - \bar{R}_1]\,(1 - \pi_1(a^1))$

$= 0 + 0.5\,[1-1]\,(1-\frac{1}{2}) = 0$

$H_2(a^2) = H_1(a^2) - \alpha\,[R_1 - \bar{R}_1]\,\pi_1(a^2)$

$= 0 - 0.5\,[1-1]\,\frac{1}{2} = 0$

(2) $H_2(a^1) = 0,\ H_2(a^2) = 0$

$\pi_2(a^1) = \frac{e^0}{e^0+e^0} = \frac{1}{2}$

$\pi_2(a^2) = \frac{e^0}{e^0+e^0} = \frac{1}{2}$

$\rightarrow a_2 = a^2,\ R_2 = 0.5 \Rightarrow \bar{R}_2 = \frac{1+0.5}{2} = 0.75$

$H_3(a^1) = H_2(a^1) - \alpha\,[R_2 - \bar{R}_2]\,\pi_2(a^1)$

$= 0 - 0.5\,[0.5 - 0.75]\,\frac{1}{2} = 0.0625$

$H_3(a^2) = H_2(a^2) + \alpha\,(R_2 - \bar{R}_2)\,(1 - \pi_2(a^2))$

$= 0 + 0.5\,(0.5 - .75)\,(1 - \frac{1}{2}) = -0.0625$

③

$H_3(a^1) = 0.0625 \quad H_3(a^2) = -0.0625$

$$\pi_3(a^1) = \frac{e^{0.0625}}{e^{0.0625} + e^{-0.0625}} = 0.531$$

$$\pi_3(a^2) = \frac{e^{-0.0625}}{e^{0.0625} + e^{-0.0625}} = 0.469$$

$\rightarrow a_3 = a^1 \rightarrow R_3 = 0 \Rightarrow \bar{R}_3 = \dfrac{1+0.5+0}{3} = 0.5$

$$H_4(a^1) = H_3(a^1) + \alpha\,[R_3 - \bar{R}_3]\,(1_{a^1 = a_3} - \pi_3(a^1))$$

$$= 0.0625 + 0.5(0 - 0.5)\,(1 - 0.531) = -0.0547$$

$$H_4(a^2) = H_3(a^2) - \alpha\,[R_3 - \bar{R}_3]\,\pi_3(a^2)$$

$$= -0.0625 - 0.5\,[0 - 0.5]\,0.469 = 0.0547$$

---

④ $H_4(a^1) = -0.0547, \quad H_4(a^2) = 0.0547$

$$\pi_4(a^1) = \frac{e^{-0.0547}}{e^{-0.0547} + e^{0.0547}} = 0.473$$

$$\pi_4(a^2) = \frac{e^{0.0703}}{e^{-0.0547} + e^{0.0547}} = 0.527$$