1 Chapter 1

2 Chapter 2

Exercise 2.1. *In "ϵ-greedy action selection, for the case of two actions and " $\epsilon = 0.5$, what is the probability that the greedy action is selected?*

Write $G :=$ {greedy action is selected}. We have

$$\begin{aligned}\mathbb{P}[\,\mathrm{G}\,] &= \mathbb{P}[\,\mathrm{G}\,, random] + \mathbb{P}[\,\mathrm{G}\,, optimal] \\ &= \epsilon\frac{1}{2} + (1-\epsilon) \\ &= 0.75.\end{aligned}$$

Exercise 2.2. *Bandit example Consider a k-armed bandit problem with $k = 4$ actions, denoted $1, 2, 3,$ and 4. Consider applying to this problem a bandit algorithm using ε-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A1 = 1, R1 = -1, A2 = 2, R2 = 1, A3 = 2, R3 = -2, A4 = 2, R4 = 2, A5 = 3, R5 = 0$. On some of these time steps the ε case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?*

We denote the sample average action value after n-steps as

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i 1_{A_i=a}}{\sum_{i=1}^{t-1} 1_{A_i=a}}. \tag{1}$$

Timestep t	$Q_{t+1}(1)$	$Q_{t+1}(2)$	$Q_{t+1}(3)$	$Q_{t+1}(4)$	Greedy action	Action selected
t=0	0	0	0	0	-	$A_1 = 1$
t=1	$\frac{R_1}{1_{A_1=1}} = -1$	0	0	0	2, 3 or 4	$A_2 = 2$
t=2	-1	1	0	0	2	$A_3 = 2$
t=3	-1	$\frac{R_2+R_3}{2} = -0.5$	0	0	3 or 4	$A_4 = 2$
t=4	-1	$\frac{R_2+R_3+R_4}{3} = \frac{1}{3}$	0	0	3	$A_5 = 3$
t=5	-1	1/3	0	0	3	end

1. action $A_1 = 1$ can be either exploration or exploitation because the optimal value is zero for all actions.
2. action $A_2 = 2$ could be either exploration or exploitation because action=2 is optimal too.
3. action $A_3 = 2$ could be either exploration or exploitation because action=2 is optimal too.
4. action $A_4 = 2$ is exploration because the action 2 is not optimal.
5. action $A_5 = 4$ is exploitation because action=3 is the only optimal choice.

Exercise 2.3. *In the comparison shown in Figure 2.2, which method will perform best in the long run in terms of cumulative reward and probability of selecting the best action? How much better will it be? Express your answer quantitatively*

cumulative reward We have for optimal action a_*

$$\mathbb{E}[R_i|A_i = a_*] = q_*(a_*). \tag{2}$$

The probability of choosing the optimal action among n-options is

$$\mathbb{P}[A_i = a_*] = (1-\varepsilon) + \frac{\varepsilon}{n}, \tag{3}$$

and $\mathbb{P}[exploration] = 1 - \mathbb{P}[A_i = a_*]$. The true value $q_*(a)$ each of the ten actions $a = 1, ..., 10$ was selected according to a normal distribution with mean zero and unit variance and so

$$\mathbb{E}[\mathbb{E}[R_i|exploration]] = \mathbb{E}\left[\sum_{action\ a} \mathbb{E}[R_i|A_i = a]\right] = \sum_{action\ a} \mathbb{E}[q_*(a)] = \sum_{action\ a} 0 = 0. \tag{4}$$

So the cumulative reward is from the law of total expectation

$$\begin{aligned}\mathbb{E}[R_i] =& \mathbb{E}[\mathbb{E}[R_i|A_i = a_*]]\mathbb{P}[A_i = a_*] + \mathbb{E}[\mathbb{E}[R_i|exploration]]\mathbb{P}[exploration] \\ =& q_*(a_*)((1-\varepsilon) + \frac{\varepsilon}{n}) + 0(\varepsilon - \frac{\varepsilon}{n}) \\ =& q_*(a_*)((1-\varepsilon) + \frac{\varepsilon}{n}).\end{aligned}$$

We have

$$\mathbb{E}_{0.01}[R_i] = q_*(a_*)0.991 > q_*(a_*)0.91 = \mathbb{E}_{0.1}[R_i]. \tag{5}$$

probability of selecting the best action Write $G := \{\text{greedy action is selected}\}$. We have for $\varepsilon = 0.01$

$$\begin{aligned}\mathbb{P}_{0.01}(\text{ G }) =& \mathbb{P}(\text{ G }, random) + \mathbb{P}(\text{ G }, optimal) \\ =& \epsilon\frac{1}{k} + (1-\epsilon) \\ =& 0.01\frac{1}{10} + 0.99 = 0.991\end{aligned}$$

and similarly for $\varepsilon = 0.1$

$$\begin{aligned}Prob_{0.1}(\text{ G }) =& Prob(\text{ G }, random) + Prob(\text{ G }, optimal) \\ =& \epsilon\frac{1}{k} + (1-\epsilon) \\ =& 0.1\frac{1}{10} + 0.9 = 0.91\end{aligned}$$

So we see that the first probability is slightly higher than the second one.

Exercise 2.4.

Exercise 2.5.

Exercise 2.6.

Exercise 2.7.