Chapter 2 Solutions

Reinforcement Learning: An Introduction

Alireza Azimi

2026-05-24

Table of contents

1 Exercise 2.1

$$\begin{aligned} P(A_t = \arg\max_a Q(a)) &= P(A_t = \arg\max_a Q(a) | 1-\epsilon) \times (1-\epsilon) \\ &\quad + P(A_t = \arg\max_a Q(a) | \epsilon) \times \epsilon \\ &= 1 \times (1-\epsilon) + \frac{1}{2} \times \epsilon \\ &= 0.5 + 0.25 = 0.75 \end{aligned}$$

2 Exercise 2.2

Let's step through each time step and see

Time step 1 $A_1 = 1, Q_1(1) = -1$ and the rest are 0.

Time step 2 $A_2 = 2, Q_2(2) = 1, Q_2(1) = -1$ and the rest are 0.

Time step 3 $A_3 = 2, Q_3(2) = \frac{-1}{2}, Q_3(1) = -1$ and the rest are 0.

Time step 4 $A_4 = 2, Q_4(2) = \frac{1}{3}, Q_4(1) = -1$ and the rest are 0.

Time step 5 $A_5 = 3, Q_5(2) = \frac{1}{3}, Q_5(1) = -1$ and the rest are 0.

A random action selection **definitely occured at time steps 4 and 5**.

A random action selection **possibly occured on time steps 1, 2 and 3**.

3 Exercise 2.3

In the long run as the number of time-steps approaches infinity. That means for methods which $\epsilon > 0$ the

action value would approach it's true mean value due to the law of large numbers. i.e. $Q_t(a) \to Q(a)$ as $t \to \infty$.

For $\epsilon = 0$:

$$0.2 \leq P(optimal) \leq 0.4$$

For $\epsilon = 0.01$ as:

$$P(optimal) = \epsilon * 0.1 + (1 - \epsilon) = 0.01 * 0.1 + 0.99 = 0.991$$

For $\epsilon = 0.1$

$$P(optimal) = \epsilon * 0.1 + (1 - \epsilon) = 0.1 * 0.1 + 0.9 = 0.91$$

We expect $\epsilon = 0.01$ to outperform the other configurations in the long-run.

4 Exercise 2.4

$$
\begin{aligned}
Q_{n+1} &= Q_n + \alpha_n[R_n - Q_n] \\
&= Q_n + \alpha_n R_n - \alpha_n Q_n \\
&= (1-\alpha_n)Q_n + \alpha_n R_n \\
&= (1-\alpha_n)(Q_{n-1} + \alpha_{n-1}[R_{n-1} - Q_{n-1}]) + \alpha_n R_n \\
&= (1-\alpha_n)Q_{n-1} + (1-\alpha_n)\alpha_{n-1}R_{n-1} - (1-\alpha_n)\alpha_{n-1}Q_{n-1} + \alpha_n R_n \\
&= (1-\alpha_n)(1-\alpha_{n-1})Q_{n-1} + (1-\alpha_n)\alpha_{n-1}R_{n-1} + \alpha_n R_n \\
&= \prod_{i=0}^{n-1}(1-\alpha_{n-i})Q_1 + \sum_{i=1}^{n}\alpha_i \prod_{j=i+1}^{n}(1-\alpha_j)R_i
\end{aligned}
$$

5 Exercise 2.5

Coming soon

6 Exercise 2.6

With optimistic initialization the agent explores more in the beginning as it tries more actions. Thus leading to more spikes due to having more volatility in the action values. Whereas, in the ϵ greedy method we get far lass exploration in the beginning and therefore a slower learning rate and less volatility.

7 Exercise 2.7

From Exercise 2.4 we know that:

$$
Q_{n+1} = \prod_{i=0}^{n-1}(1-\beta_{n-i})Q_1 + \sum_{i=1}^{n}\beta_i \prod_{j=i+1}^{n}(1-\beta_j)R_i
$$

Now let's expand the first term with $\beta_n = \frac{\alpha}{\overline{O}_n}$ and $\overline{O}_n = \overline{O}_{n-1} + \alpha(1-\overline{O}_{n-1})$:

$$
\begin{aligned}
\prod_{i=0}^{n-1}(1-\beta_{n-i})Q_1 &= (1-\beta_n)(1-\beta_{n-1})...(1-\beta_1)Q_1 \\
&= (\frac{\overline{O}_n - \alpha}{\overline{O}_n})(\frac{\overline{O}_{n-1} - \alpha}{\overline{O}_{n-1}})...(\frac{\overline{O}_1 - \alpha}{\overline{O}_1})Q_1 \\
&= (\frac{(1-\alpha)\overline{O}_{n-1}}{\overline{O}_n})(\frac{(1-\alpha)\overline{O}_{n-2}}{\overline{O}_{n-1}})...(\frac{(1-\alpha)\overline{O}_0}{\overline{O}_1}) \\
&= 0 \text{ because } \overline{O}_0 \text{ is } 0
\end{aligned}
$$

Q.E.D

$$Q_{n+1} = \sum_{i=1}^{n} \beta_i \prod_{j=i+1}^{n} (1 - \beta_j) R_i$$

8 Exercise 2.8

It is because the uncertainty of UCB term for each action follows $UCB \to \infty$ as $N(a) \to 0$. Since we have 10 actions (10 arms) the agent will explore all 10 actions in the first 10 steps, resulting in selecting the optimal action early on during training thus resulting in a spike in step 11. Once all actions have been tried the UCB term reduces significantly since now we have $N(a) = 1$ for all actions.

9 Exercise 2.9

Let's recall the sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Now for two actions a_1 or a_2 the softmax distribution is as follows:

$$\begin{aligned} Pr\{A_t = a\} &= \frac{e^{H_t(a)}}{e^{H_t(a)} + e^{H_t(b)}} \\ &= \frac{e^{H_t(a)}}{e^{H_t(a)}(1 + e^{H_t(b) - H_t(a)})} \\ &= \frac{1}{1 + e^{-(H_t(a) - H_t(b))}} \end{aligned}$$

10 Exercise 2.10

Coming soon