Chapter 2 - Solutions

Alessio Caciagli

March 2021

1 Exercise 2.1

In ϵ-greedy action selection, greedy action is *always* selected with probability (1-ϵ). On the other side, with probability ϵ a random selection from all the actions is performed with equal probability i.e. we assume a uniform distribution for the probability of selecting a certain action a. These two cases are of course mutually exclusive. In formal terms, assuming N actions are available, the probability of selecting the best action is:

$$p(x) = (1 - \epsilon) + \frac{\epsilon}{N} \tag{1}$$

The second term arises since $p(x = k) = 1/N$ for a uniform distribution. With $N = 2$ and $\epsilon = 0.5$, the result is $p(x) = 0.5 + 0.5 * 0.5 = 0.75$. $\square$

2 Exercise 2.2

Let's tabulate the mean rewards, $Q_t(a)$, at each time step:

Time	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$
$t = 0$	0	0	0	0
$t = 1$	1	0	0	0
$t = 2$	1	1	0	0
$t = 3$	1	3/2	0	0
$t = 4$	1	5/3	0	0
$t = 5$	1	5/3	0	0

Given this, we can reconstruct whether the ϵ case might have occurred at each time step:

- At $t = 1$, either ϵ or greedy selection might have occurred (because all actions are greedy at the start)
- At $t = 2$, ϵ selection has occurred (the greedy action is 1)
- At $t = 3$, either ϵ or greedy selection might have occurred (because both actions 0 and 1 are greedy)

- At $t = 4$, either ϵ or greedy selection might have occurred (although the greedy action 2 has been selected)

- At $t = 5$, ϵ selection has occurred (the greedy action is 2)

□

3 Exercise 2.3

Assuming both methods have converged so that the optimal action corresponds to the greedy action (which is reasonable in the limit $t \to \infty$), we have the following by applying Equation 1:

- For the $\epsilon = 0.1$ case, the probability of optimal action selection is $p(x) = 0.9 + 0.1 * 0.1 = 0.91$

- For the $\epsilon = 0.01$ case, the probability of optimal action selection is $p(x) = 0.99 + 0.01 * 0.1 = 0.991$

Since we can disregard the transient in the long run, the method with the highest probability of optimal action selection will also yield the highest cumulative reward. Hence, method $\epsilon = 0.01$ will perform best. □

4 Exercise 2.4

The estimate Q_{n+1} is given by the general formula:

$$Q_{n+1} = Q_n + \alpha_n(R_n - Q_n) \tag{2}$$

We can expand recursively, such that, for the first two expansions:

$$\begin{aligned} Q_{n+1} &= \alpha_n R_n + (1 - \alpha_n)Q_n \\ &= \alpha_n R_n + (1 - \alpha_n)[Q_{n-1} + \alpha_{n-1}(R_{n-1} - Q_{n-1})] \\ &= \alpha_n R_n + (1 - \alpha_n)[\alpha_{n-1}R_{n-1} + (1 - \alpha_{n-1})Q_{n-1}] \\ &= \alpha_n R_n + (1 - \alpha_n)\alpha_{n-1}R_{n-1} + (1 - \alpha_n)(1 - \alpha_{n-1})Q_{n-1} \end{aligned} \tag{3}$$

We can refactor it in the final form:

$$Q_{n+1} = \left(\prod_{i=1}^{n}(1 - \alpha_i)\right) Q_1 + \sum_{i=1}^{n} \alpha_i R_i \prod_{j=i+1}^{n} (1 - \alpha_j) \tag{4}$$

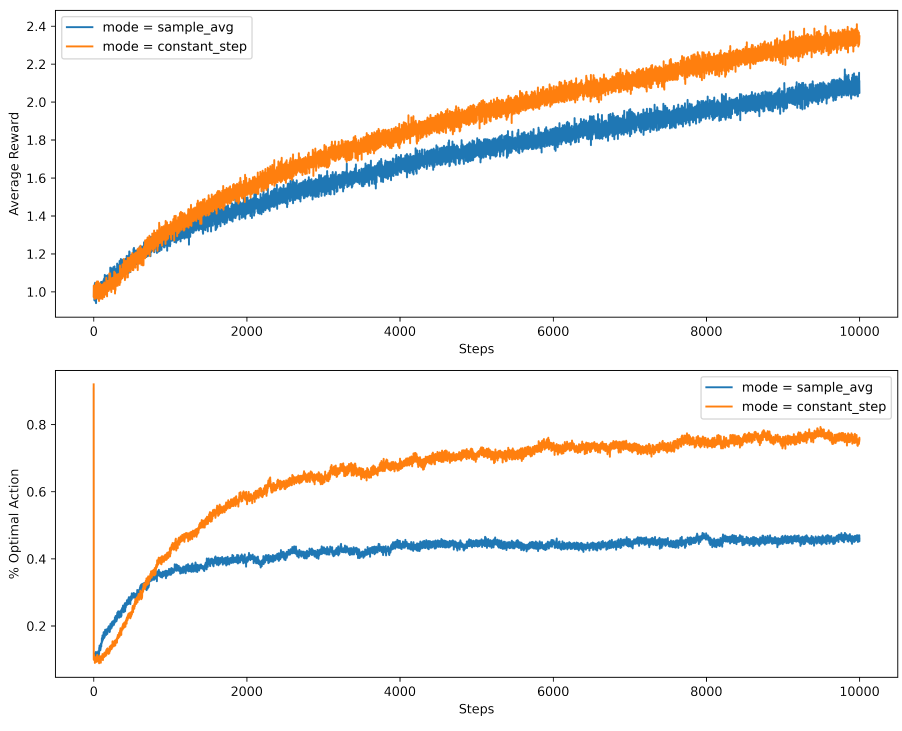
As a check, we can assume a stationary problem with weights $\alpha_i = 1/i$. In this case, the first term of Equation 4 is zero, and the second term becomes:

$$
\begin{aligned}
Q_{n+1} &= \sum_{i=1}^{n} \frac{R_i}{i} \left(1 - \frac{1}{i+1}\right) \cdots \left(1 - \frac{1}{n}\right) \\
&= \sum_{i=1}^{n} \frac{R_i}{i} \frac{i}{i+1} \cdots \frac{n-1}{n} \\
&= \frac{1}{n} \sum_{i=1}^{n} R_i
\end{aligned} \tag{5}
$$

which is the estimate of the action value for a stationary problem. Hence, Equation 4 is the estimate of the action value for the general case of step-size parameters α_n (either stationary or non-stationary). $\square$

5 Exercise 2.5

See the companion *code* folder (for code and notebooks) for the implementation.



The constant step-size method is superior to the incremental sample average method for non-stationary problems.