Chenghao Wang
CS5180 Reinforcement Learning (Fall 2022)
September 22, 2022

Ex1

Question 1. 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 Q1(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?

	A1	A2	A3	A4
$Q_1(a)$	0	0	0	0
$Q_2(a)$	-1	0	0	0
$Q_3(a)$	-1	1	0	0
$Q_4(a)$	-1	-1/2	0	0
$Q_5(a)$	-1	1/3	0	0
$Q_6(a)$	-1	1/3	0	0

T=1: greed or non-greed
T=2: greed or non-greed
T=3: greed or non-greed
T=4: non-greed
T=5: non-greed

Question 2. If the step-size parameters, a_n, are not constant, then the estimate Q_n is a weighted average of previously received rewards with a weighting different from that given by Equation 2.6. What is the weighting on each prior reward for the general case, analogous to Equation 2.6, in terms of the sequence of step-size parameters?

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

Question 3. Bias in Q-value estimates.
(a) Consider the sample-average estimate in Equation 2.1. Is it biased or unbiased? Justify your answer with brief words and equations.

In general, the sample-average estimate is a unbiased approach because that the

$$Q_t(a) = \frac{sum of rewards when a taken prior to t}{number of times a taken prior to t} = \frac{\sum_{t=1}^{t-1} R_i \cdot \mathbb{I}_{Ai=a}}{\sum_{t=1}^{t-1} \mathbb{I}_{Ai=a}} = \mathbb{E}[R_t | A_t = a] = q^*(a).$$

(b) If $Q_1 = 0$, is Q_n (for $n > 1$) biased? Justify your answer with brief words and equations.1

(b) equation 2.6

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

when $Q_1 = 0$

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

$$\Rightarrow E(Q_{n+1}) = E\left(\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i\right)$$

$$= \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} E(R_i)$$

$\Rightarrow$ when $\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$

$$E(Q_{n+1}) = E(R_i)$$

$\Rightarrow$ equation is unbiased

FIGURE 1

(c)Derive condition(s) for Q_1 for when Q_n will be unbiased.2

(c) Q_n will be unbiased if $\begin{cases} Q_1 = 0 \\ \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1 \end{cases}$

FIGURE 2

(d)Show that Q_n is an unbiased estimator as nß (which is often referred to as asymptotically unbiased).3

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

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

$$= \lim_{n\to\infty} \alpha \cdot \frac{1-(1-\alpha)^{n-1}}{1-(1-\alpha)} \cdot R_i$$

$$= \lim_{n\to\infty} \alpha \frac{1-(1-\alpha)^{n-1}}{\alpha} \cdot R_i$$

$$= \lim_{n\to\infty} 1-(1-\alpha)^{n-1} \cdot R_i$$

$$= R_i$$

$$\Rightarrow \lim_{n\to\infty} E(Q_{n+1}) = E(R_i)$$

FIGURE 3

(e) Why should we expect that the exponential recency-weighted average will be biased in practice? Think about what happens to Q_1 or a in practice.

For a, we are able to control the learning rate by modify a, the learning process could be speed up by increasing a and vice versa. For Q_1, the agent could start explore more in the beginning rather then always choosing the best choice by increase Q_1.

Question 4. Show that in the case of two actions, the soft-max distribution is the same as that given by the logistic, or sigmoid, function often used in statistics and artificial neural networks.4

4. Softmax: $$\pi_a = \frac{\exp\{H_t(a)\}}{\sum_{b=1}^{k} \exp\{H_t(b)\}}$$

case of two action:

$$\pi_a = \frac{\exp\{H_t(a)\}}{\exp\{H_t(a)\} + \exp\{H_t(b)\}} = \frac{1}{1 + \exp\{H_t(b) - H_t(a)\}}$$

if $\theta = -(H_t(b) - H_t(a))$

then $$\pi_a = \frac{1}{1 + \exp\{-\theta\}} = \sigma(\theta)$$ sigmoid

FIGURE 4

Question 5. Implement the ε-greedy algorithm with incremental updates. Note that in the graph: "All the methods formed their action-value estimates using the sample-average technique (with an initial estimate of 0). See equation 2.1"

From figures below we can notice that the average reward getting larger with higher ε value and smaller with lower ε value56, the reason behind this result is that the higher ε, the more exploration for the agent so that the higher chance to reaching better action.

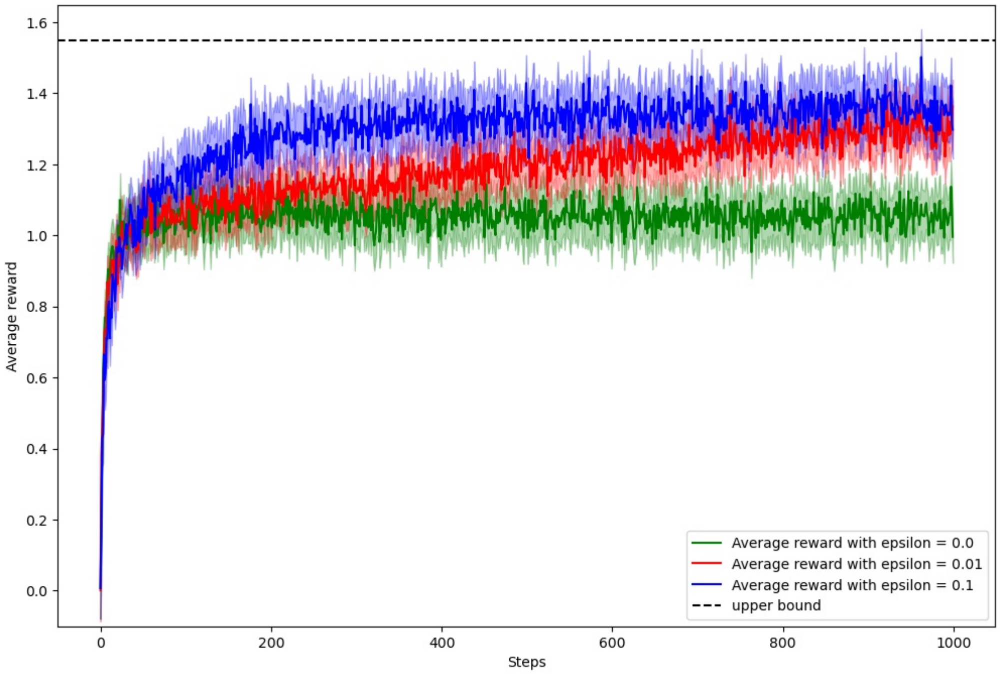

FIGURE 5. Average reward

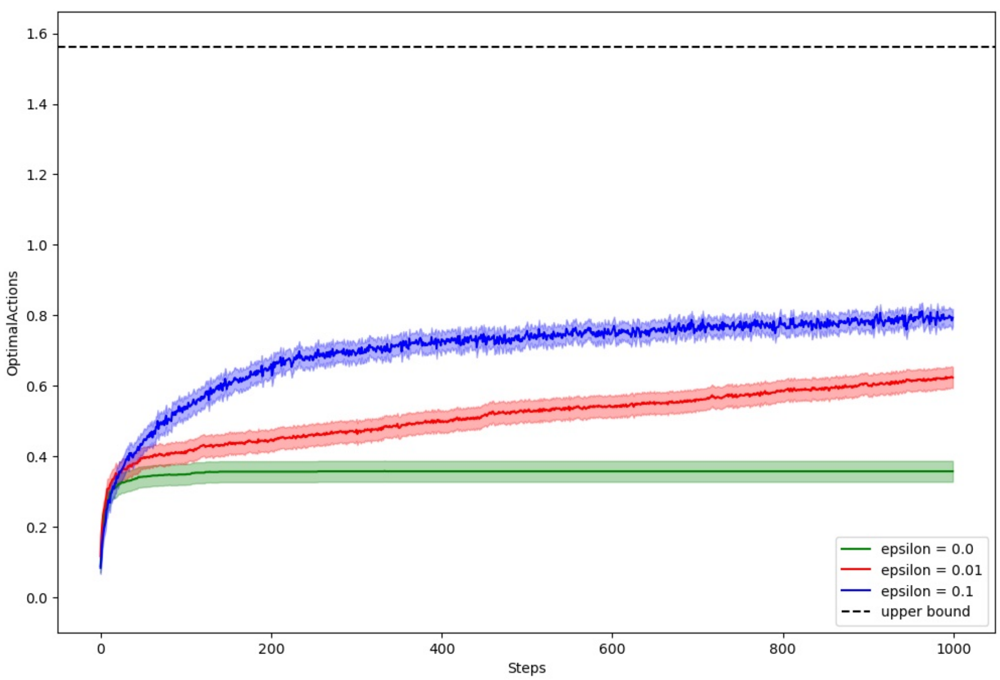

FIGURE 6. Optimal Action

Question 6. We have seen the spike for optimistic initialization in class (Figure 2.3, Figure 2.4 in RL2e). Observe that UCB also produce spikes in the very beginning in both the two reproduced figures. Explain in your own words why the spikes appear (both the sharp increase and sharp decrease). Analyze and use your experimental data as further empirical evidence to back your reasoning.

In the optimistic initialization, the first few steps are not really random. Instead, we loop through all the actions multiple times, picking our large optimistic value at random. In the first round, all actions have equal optimization percentages (10%).78 The next action will have a spike, because on average the best action will have the largest value. This will be repeated with decreasing values until the effect of the optimal initial value fades away.

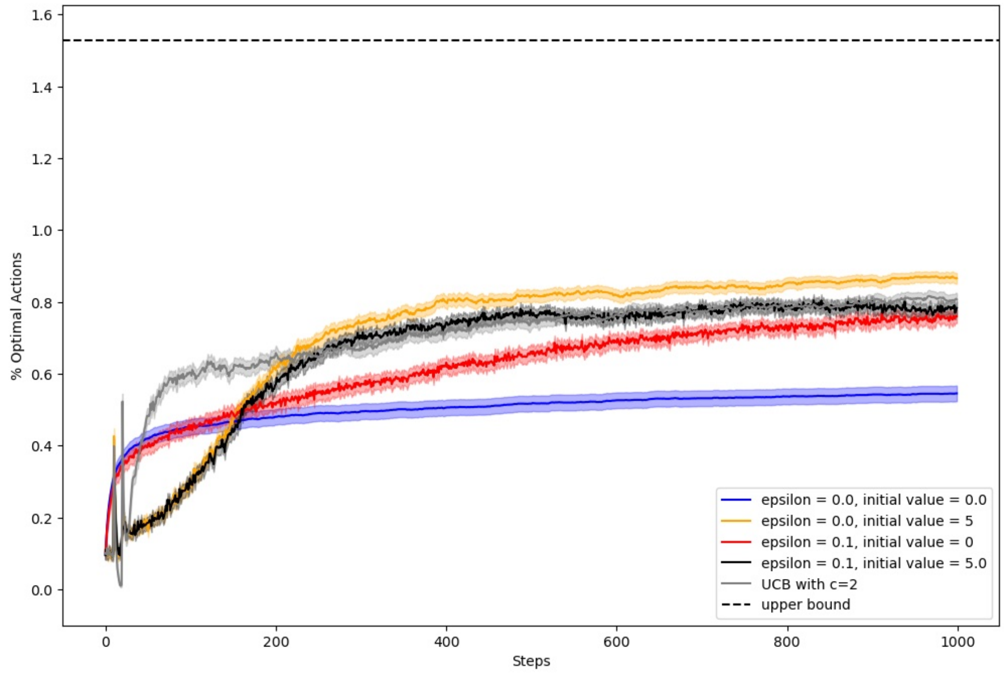

FIGURE 7. Fig2.3

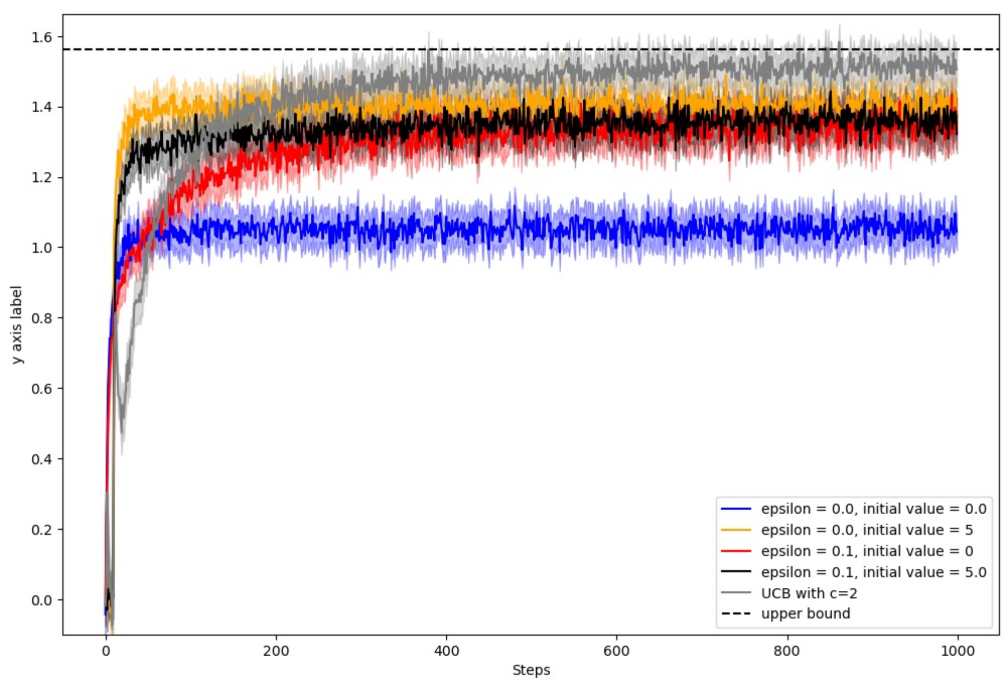

FIGURE 8. Fig2.4