'''

CS 5180 Fall 2022

Exercise 1: Multi-armed Bandits

Hongyan Yang

'''

P1 Written (RL2e 2.2) Exploration vs. exploitation:

Based on the information given.

At time step 4 and 5 the ε case definitely have occurred. At step 4, Q4(2) = -0.5 < Q4(3) and Q4(4), while A4 = 2 and it is not a greedy action. At step 5, Q5(3) = 0 < Q5(2), while A5 = 3 and it is not a greedy action. So at time step 4 and 5 the ε case definitely have occurred.

At time step 1, 2 and 3 the ε case could possibly have occurred. At step 1, Q1(1) = 0 = argmax a Q1(a), and A1 = 1 which might be a ε case. At step 2, Q2(2) = 0 = argmax a Q2(a), and A2 = 2 which might be a ε case. At step 3, Q3(2) = 1 = argmax a Q3(a), and A3 = 2 which might be a ε case. So at time step 1, 2 and 3 the ε case could possibly have occurred.

P2 Written (RL2e 2.4) Varying step-size weights:

2. Based on the information given,

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

P3 Bias in Q-value estimates:

3(a). The sample-average estimate in E2.1 is unbiased

Because by nature $Q_t(a) = \hat{q}^*_a$,

$$\text{bias}(\hat{q}^*_a) = E(\hat{q}^*_a) - q^*_a$$

$$= E\left[\frac{1}{m}\sum_{i=1}^{m} Q_i(a)\right] - q^*_a$$

$$= \frac{1}{m}\sum_{i=1}^{m} E(Q_i(a)) - q^*_a$$

$$= \frac{1}{m} \times m \times q^*_a - q^*_a$$ $\therefore$ it is an unbiased estimate.

$$= q^*_a - q^*_a = 0$$

3(b). If $Q_1=0$, Q_{n+1} (for $n>1$) $= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$

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

$\therefore \hat{q^*} = E(Q_n) = E[\alpha \sum_{i=1}^{n} (1-\alpha)^{n-i} R_i]$

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

$= \alpha \times \dfrac{1-(1-\alpha)^n}{\alpha} E(R_i)$

$= [1-(1-\alpha)^n] q^*$

$\therefore bias(\hat{q^*}) = E(Q_{n+1}) - q^* = [1-(1-\alpha)^n] q^* - q^* = [-(1-\alpha)^n] q^* \neq 0$

$\therefore Q_n$ (for $n>1$) is a biased estimate for q^*.

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

Based on solution to 3(b), $\hat{q^*} = (1-\alpha)^n E(Q_1) + [1-(1-\alpha)^n] q^*$

$\therefore bias(\hat{q^*}) = E(Q_{n+1}) - q^* = (1-\alpha)^n (E(Q_1) - q^*)$

For Q_n to be an unbiased estimate for q^*,

$bias(\hat{q^*}) = (1-\alpha)^n (E(Q_1) - q^*) = 0$

$\therefore$ When $Q_1 = q^*$, Q_n will be unbiased

(d). Based on the solutions in 3(b) and 3(c)

$bias(\hat{q^*}) = E(Q_{n+1}) - q^* = (1-\alpha)^n [E(Q_1) - q^*]$

$\because 0<\alpha<1 \Rightarrow 0<1-\alpha<1$, $\therefore \lim_{n\to+\infty} bias(\hat{q^*}) = \lim_{n\to+\infty} (1-\alpha)^n [E(Q_1) - q^*]$

$\lim_{n\to+\infty} (1-\alpha)^n = 0$

$= 0 \times [E(Q_1) - q^*] = 0$

$\therefore Q_n$ is an unbiased estimator as $n\to\infty$

(e) Exponential recency-weighted average will be biased because:

① In practice we don't know the true q^* and it's impractical to set $Q_1 = q^*$ to let Q_n to be unbiased.

② In practice we always choose $\alpha \in (0,1)$ and a relatively small α such as 0.1 but not 1. So $bias(\hat{q^*}) = (1-\alpha)^n (E(Q_1 - q^*)$ will not be 0 to let Q_n to be unbiased.

4. (RL2e 2.9) Gradient Bandit:

4 According to the information given,

$$Pr\{A_t = a\} = \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}}$$

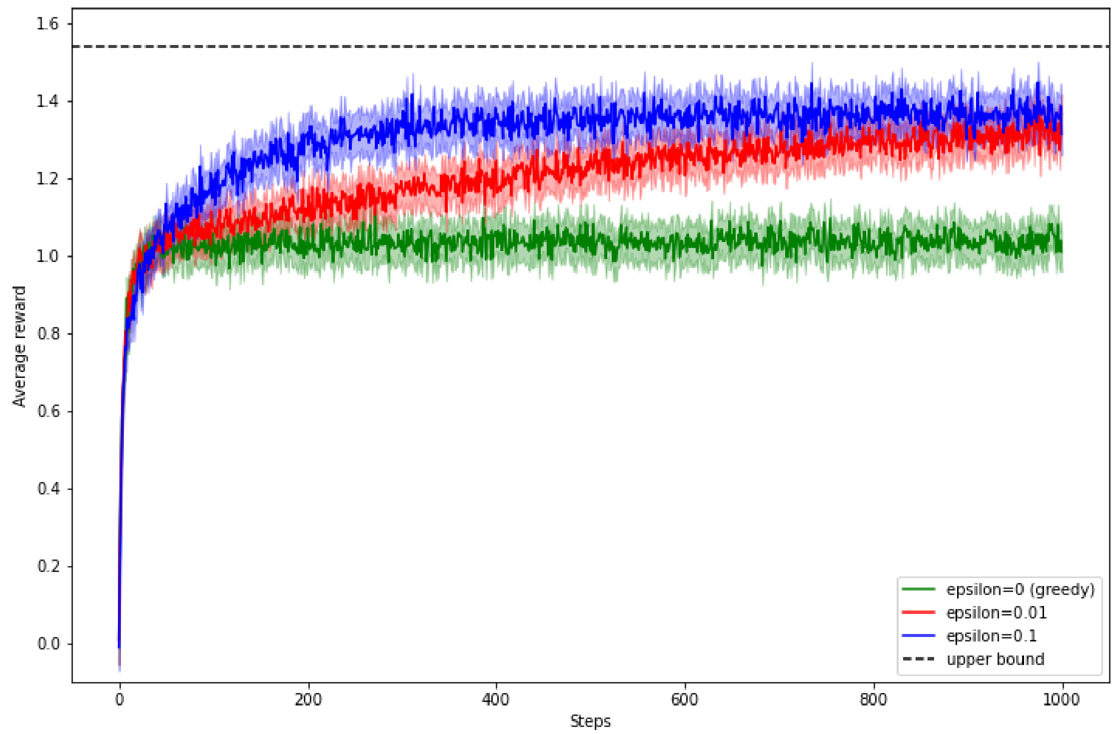
In the case of two actions $Pr\{A_t = 1\} = \frac{e^{H_t 1}}{e^{H_t 1} + e^{H_t 2}} = \frac{1}{1 + e^{-(H_t 1 - H_t 2)}}$

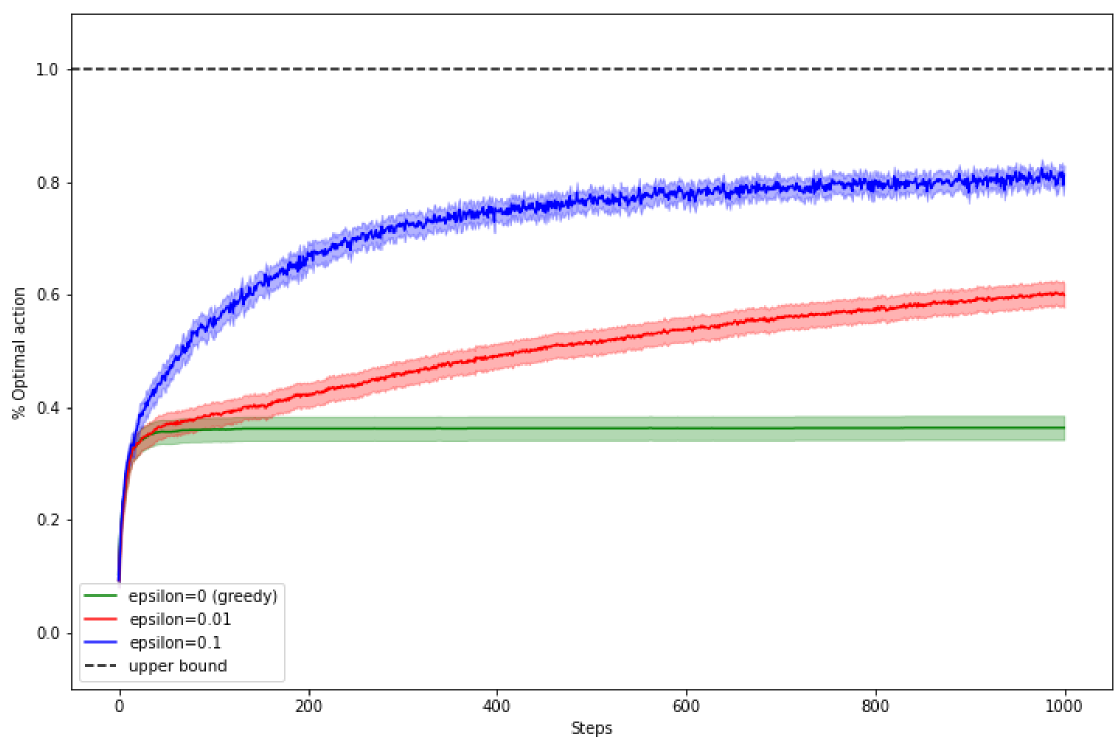
$$Pr\{A_t = 2\} = \frac{e^{H_t 2}}{e^{H_t 1} + e^{H_t 2}} = \frac{1}{1 + e^{-(H_t 2 - H_t 1)}}$$

∴ The distribution is the same as Sigmoid function.

5. Reproducing Figure 2.2. (RL2e page 29):

5. Plot:





5. Written:

1). The average rewards that the algorithm converges to using different ϵ values:

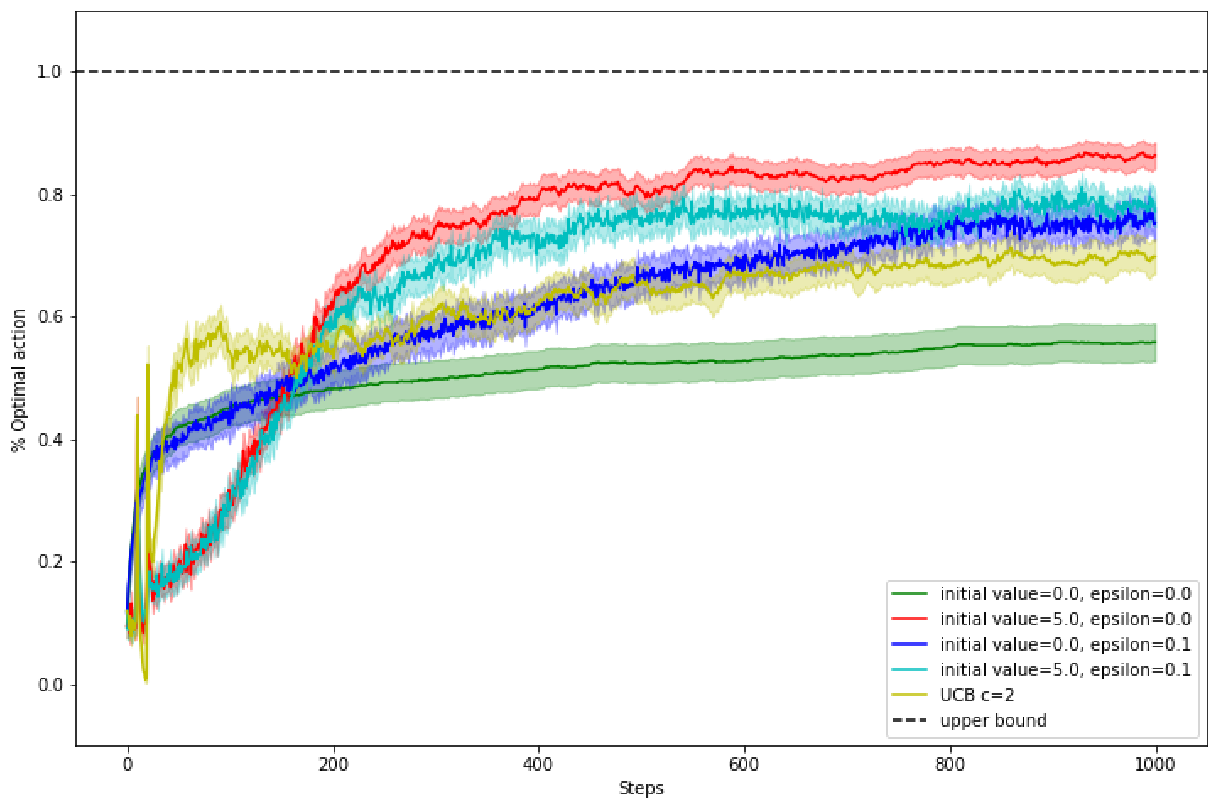
Epsilon=0: The algorithm will finally get stuck performing suboptimal actions at one of the arms and get the average reward of that specific arm.

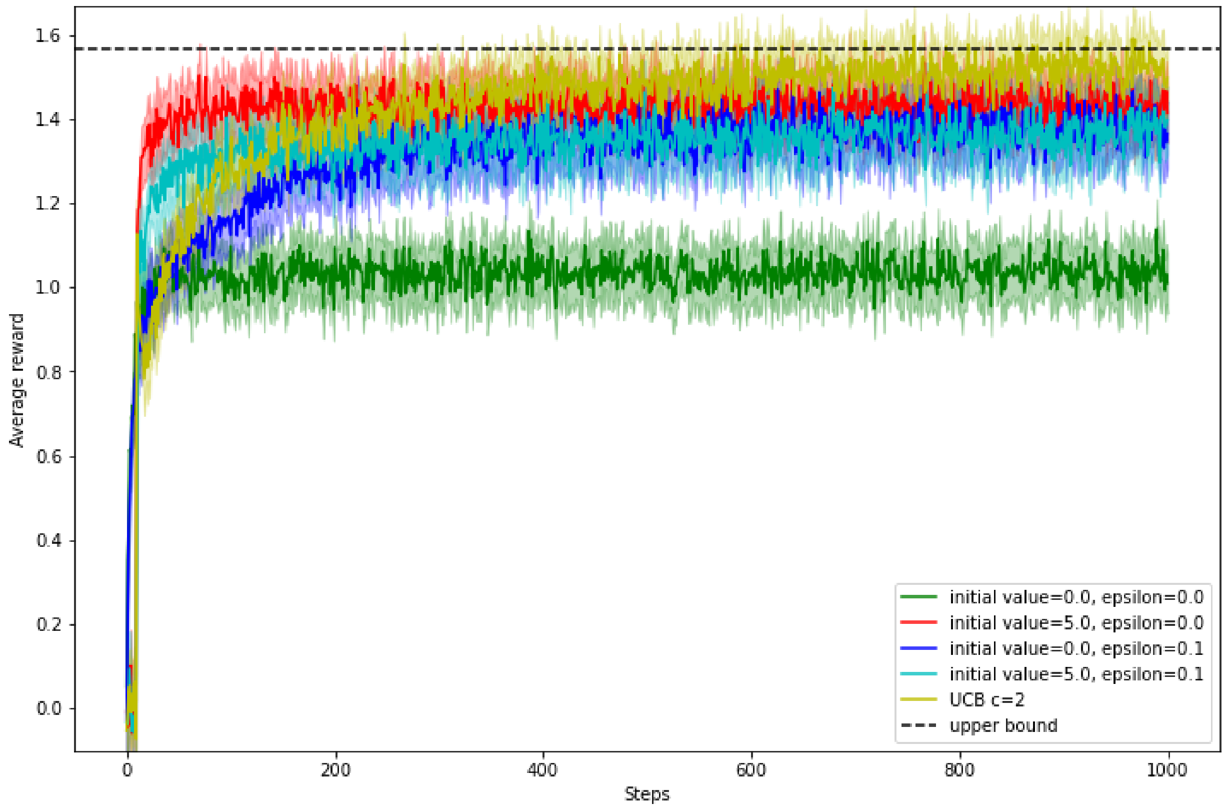
Epsilon=0.01: 0.99 * ($\max_a q_*(a)$) + 0.01 * average($q_*(a)$)

Epsilon=0.1: 0.9 * ($\max_a q_*(a)$) + 0.1 * average($q_*(a)$)

6. Reproducing and supplementing Figures 2.3 (RL2e page 34) and 2.4. (RL2e page 36):

6. Plot:





6. Written:

Why the spikes appear:

1). Sharp increase:

For optimistic initialization:

At the very beginning steps, with an optimistic Q0 = 5, to start with, which is much larger than the true q* = 0. Every arm is very likely to be the “optimal” arm in their first several steps because their Q have not been pulled down by their true q* yet, so does the true optimal arm. It will remain to be the optimal action until it’s Q* is below 5. So comes the sharp increase.

For UCB:

Before the optimal arm was selected, it’s Nt(a) = 0, and it will be surely be selected in the very beginning by algorithm and likely to produce a higher reward to pull up the average reward sharply. That’s why it comes the sharp increase.

2). Sharp decrease:

For optimistic initialization:

Because the true q* = 0, much smaller than Q0, which is 5. After the true optimal arm selected to be the optimal action, its value will quickly be reduced to around 0 after several steps. And other arms will be selected with their Q0 = 5. So comes the sharp decrease.

For UCB:

Before every arm has been selected once. Their Nt(a) = 0, and all arms will be selected at least once in the very beginning by the design of algorithm. After the optimal arm is pulled when theirs other arms haven't been pulled, one of the other arm will surely be pulled in the next step. That's why it shows a sharp decrease.