Assignement: ex1

Ian Steenstra

September 2024

1 Problem 1

Definitely Explored: Time Step 1 and Time Step 5.
Possibly Explored: Time Steps 2, 3, and 4.

2 Problem 2

$$\text{Weighting of } R_j = \alpha_j \cdot \prod_{i=j+1}^{n} (1 - \alpha_i)$$

where:

1. α_j is the step-size parameter used at time step j when reward R_j was received.
2. $\prod_{i=j+1}^{n}(1 - \alpha_i)$ is the product of $(1 - \alpha_i)$ for all step-size parameters α_i from time step $i = j + 1$ up to $i = n$ (the current time step).

3 Problem 3

1. **Sub-Problem (a)**
   (i) **Answer:** Unbiased.
   (ii) **Reasoning:** The sample average equally weights all rewards. Since rewards are drawn from a distribution with mean q^*, the expected value of the average converges to q^*.
2. **Sub-Problem (b)**
   (i) **Answer:** Biased
   (ii) **Reasoning:** With $Q_1 = 0$ and Q_n for $n < 1$, the initial zero value continues to have a diminishing but persistent influence on all future estimates, pulling them down.

3. **Sub-Problem (c)** Let us take Equation 2.6:

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

Now, let's take the expectation of both sides:

$$E[Q_n] = (1-\alpha)^k Q_1 + \sum_{i=1}^{k} \alpha(1-\alpha)^{k-i} E[R_i]$$

Since we know the expected reward is always q^* ($E[R_i] = q^*$), we can substitute:

$$E[Q_n] = (1-\alpha)^k Q_1 + \sum_{i=1}^{k} \alpha(1-\alpha)^{k-i} q^*$$

Simplify it, given that it's a geometric sequence:

$$E[Q_n] = (1-\alpha)^k Q_1 + q^*(1-(1-\alpha)^k)$$

To make Q_n unbiased, we need $E[Q_n] = q^*$. So let's set them equal and solve for Q_1:

$$q^* = (1-\alpha)^k Q_1 + q^*(1-(1-\alpha)^k)$$

After a bit of algebra, we find that:

$$Q_1 = q^*$$

Conclusion: The condition for Q_n to be unbiased for a specific n, with a constant α and non-zero Q_1, is simply that $Q_1 = q^*$.

4. **Sub-Problem (d)** As n approaches infinity, the influence of the initial estimate Q_1 becomes negligible due to the repeated multiplication by $(1-\alpha)$, making the bias disappear in the long run.

5. **Sub-Problem (e)** We typically don't know the true expected reward q^* to initialize for unbiasedness, and the weighting scheme inherently favors recent rewards over older ones, leading to bias.

4 Problem 4

In the case of two actions with preference values $[H_1, H_2]$, the softmax distribution aligns with the logistic (sigmoid) function. Specifically, the softmax probability of selecting the first action, $\pi(1) = \frac{e^{H_1}}{e^{H_1}+e^{H_2}}$, can be manipulated by dividing the numerator and denominator by e^{H_1} and applying exponent properties to yield $\pi(1) = \frac{1}{1+e^{H_2-H_1}} = \sigma(H_1 - H_2)$, where $\sigma(H)$ is the logistic function. Since $\pi(1) + \pi(2) = 1$, it follows that $\pi(2) = 1 - \sigma(H_1 - H_2)$. This demonstrates the equivalence between the two-action softmax distribution and the logistic function applied to the two action preferences.

5 Problem 5

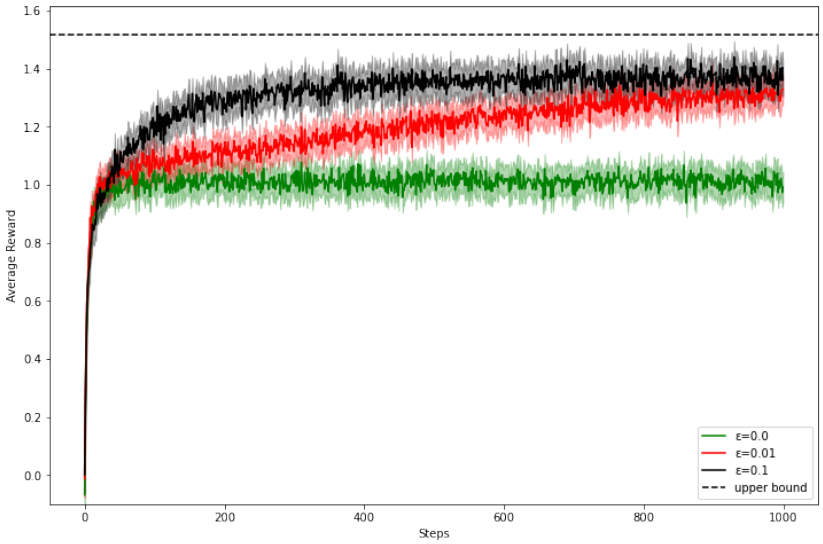

Figure 1: Problem 5: Average Reward

Written: The average rewards converge to: ≈ 1.0 for $\epsilon = 0$, ≈ 1.3 for $\epsilon = 0.01$, and ≈ 1.4 for $\epsilon = 0.1$. The reason why each converges to different average rewards is because they vary in exploration and exploitation, which directly affects their ability to find and exploit the optimal actions. As expected, the agent with a higher exploration rate ($\epsilon = 0.1$) eventually achieves the highest average reward, as its increased exploration allows it to more effectively discover and exploit the optimal actions in the long run. However, it's important to note that if ϵ were significantly higher, the agent might explore too frequently and not exploit the optimal actions often enough, leading to a lower overall reward.

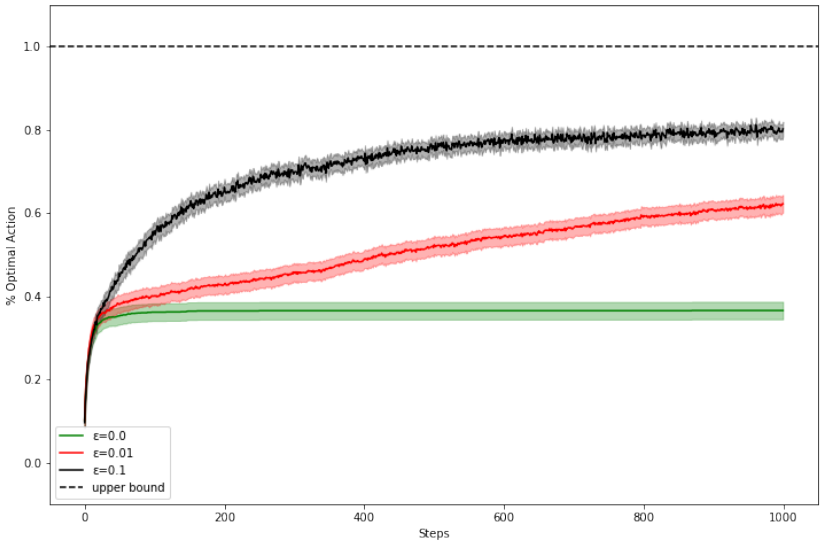

Figure 2: Problem 5: % Optimal Action

6 Problem 6

Written: Both optimistic initialization and UCB produce initial spikes in performance because they encourage high initial exploration. This leads to a sharp increase in average reward (or % optical action) as the agents quickly discover good actions. The subsequent decrease occurs as the initial exploration fades out, and the agents haven't yet converged to consistently selecting the true optimal action. The experimental data supports this, as the spikes are most prominent in configurations with higher exploration (high initial Q-values or higher UCB parameter c).

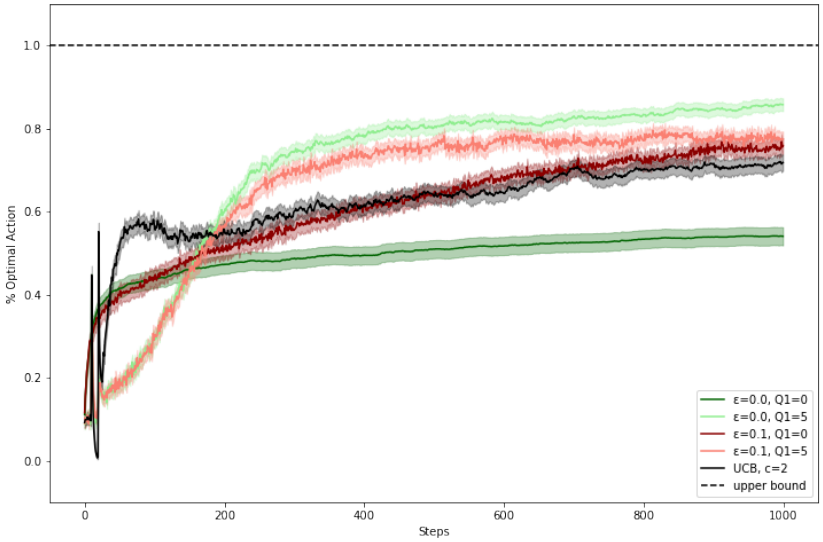

Figure 3: Problem 6: % Optimal Action

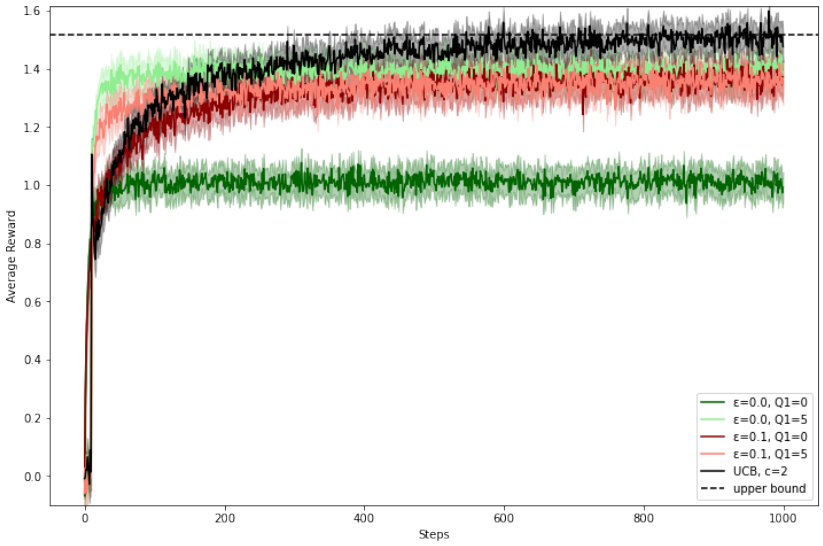

Figure 4: Problem 6: Average Reward