Ex 1 — Multi-arm Bandits — Sub: Aadhar Bansal

Wednesday, September 21, 2022 9:39 PM

5180

1. **1 point.** (RL2e 2.2) *Exploration vs. exploitation.*
Written: 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 $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. 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?

Timestep	Q_1	Q_2	Q_3	Q_4
0	0	0	0	0
1	-1	0	0	0
2	-1	1	0	0
3	-1	-0.5	0	0
4	-1	-0.5	0	2
5	-1	0.5	0	2

Action selected at Random

At $T=1$, $A_1 = 1 \rightarrow R = -1$, therefore Action must have been **random**

At $T=2$, $A_2 = 2$, $R = 1$ Most reward Greedy

At $T=3$, $A_3 = 2$, $\Sigma R = -0.5$ Selecting must be **random** as Reward = 0 is greater

At $T=4$, $A_4 = 4$, $R = 2$ Selecting Greedly

At $T=5$, $A_5 = 3$, $R = 0$ Selection must have been **random** as maximum reward not choosen

2. **1 point.** (RL2e 2.4) *Varying step-size weights.*
Written: If the step-size parameters, α_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?

$$Q_{n+1} \doteq Q_n + \alpha_n [R_n - Q_n] \quad - (2.5)$$

$$= \alpha_n R_n + (1-\alpha_n) Q_n \quad \text{expanding } Q_n \text{ as per eqn (2.5)}$$

$$= \alpha_n R_n + (1-\alpha_n)\left[Q_{n-1} + \alpha_{n-1}[R_{n-1} - Q_{n-1}]\right]$$

$$= \alpha_n R_n + \alpha_{n-1}R_{n-1} - \alpha_n\alpha_{n-1}R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})Q_{n-1}$$

$$= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1}R_{n-1} + (1-\alpha)(1-\alpha_{n-1})\alpha_{n-2}R_{n-2}$$

$$+ (1-\alpha_n)(1-\alpha_{n-1})(1-\alpha_{n-2})Q_{n-2}$$

$$\vdots$$

$$Q_{n+1} \doteq \sum_{i}^{n} \alpha_n R_n - R_{n-1}\prod_{i}^{n}\alpha_n + Q_1 \prod_{i}^{n}(1-\alpha_n)$$

Q.3

2 points. Bias in Q-value estimates.
This question is required for 5180 and extra credit for 4180
Written: Recall that $Q_n \triangleq \frac{R_1+\ldots+R_{n-1}}{n-1}$ is an estimate of the true expected reward q_* of an arbitrary arm a.
We say that an estimate is biased if the expected value of the estimate does not match the true value, i.e., $E[Q_n] \neq q_*$ (otherwise, it is unbiased).

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

$$Q_n = \frac{1}{n-1}(R_1 + R_2 \ldots R_{n-1}) \quad - \text{eq } (2.1)$$

& we know

$$E(R_n) = q_*$$

$$E(Q_n) = \frac{1}{n-1} E\left|R_1 + R_2 \ldots R_{n-1}\right|$$

$$= \frac{1}{n-1}\left|E|R_1| + E|R_2| \ldots E|R_{n-1}|\right]$$

$$= \frac{1}{n-1} \times (n-1) q_*$$

$$= q_*$$

Therefore Sample average Estimate in eq 2.1 is unbiased

For the remainder of the question, consider the exponential recency-weighted average estimate in Equation 2.5.
Assume that $0 < \alpha < 1$ (i.e., it is strictly less than 1).
(b) If $Q_1 = 0$, is Q_n (for $n > 1$) biased? Justify your answer with brief words and equations.

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

as $Q_1 = 0$

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

Expectation $E|Q_{n+1}|$

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

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

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

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

$E|Q_{n+1}| = q_*$, It will be Unbiased then

else biased

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

as per part b

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

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

$$= (1-\alpha)^n E|Q_1| + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} q_*$$

$\therefore$ It will be Unbiased when $Q_1 = 0$ & $\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$

(d) Show that Q_n is an unbiased estimator as $n \to \infty$ (which is often referred to as asymptotically unbiased).

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

$\forall$ $n \to \infty$, $0 < \alpha < 1$ $\Rightarrow$ $(1-\alpha)^n \to 0$

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

as $n \to \infty$

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

$(1-\alpha)^{n-i}$ decrease to zero, $1-(1-\alpha)^{n-1} \to 1$

i.e. $\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i}$ Converge to 1

$\therefore$ Q_n is asymptotically unbiased as $n \to \infty$.

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

As the real senario is not Ideal, we can never reach to infinite values, but sufficiently high enough, but still leads to some bias

we can Control the initial hyperparameter like Q_1, α which can control the initial environment.

Therefore biasness is more expected.

4. 1 point. (RL2e 2.9) Gradient Bandit
Written: 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.

Sigmoid: probability of selecting class / action a given $H_t(a)$ in time t

$$S_a = \frac{e^{H_t(a)}}{\sum_{b=1}^{K} e^{H_t(b)}}$$

Binary Case

$$S_a = \frac{e^{H_t(a)}}{e^{H_t(a)} + e^{H_t(b)}} = \frac{1}{1 + e^{H_t(b) - H_t(a)}} = \frac{1}{1+e^{-\theta}}$$

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

$\therefore$ Softmax distribution is same as the logistic, or sigmoid function when $\theta = H_t(a) - H_t(b)$

when $H_t(b) = 0$, both f^n returns same value for same f^n

Q.5

In the long run $\varepsilon = 0.01$ perform better

$\varepsilon = 0$, will always choose the action which discovered first & being greedy

$\varepsilon = 0.1$ = Optimal action %

$$0.9 \times 1 + 0.1 \times \frac{1}{10} = 91\%$$

$\varepsilon = 0.01$ = optimal action %

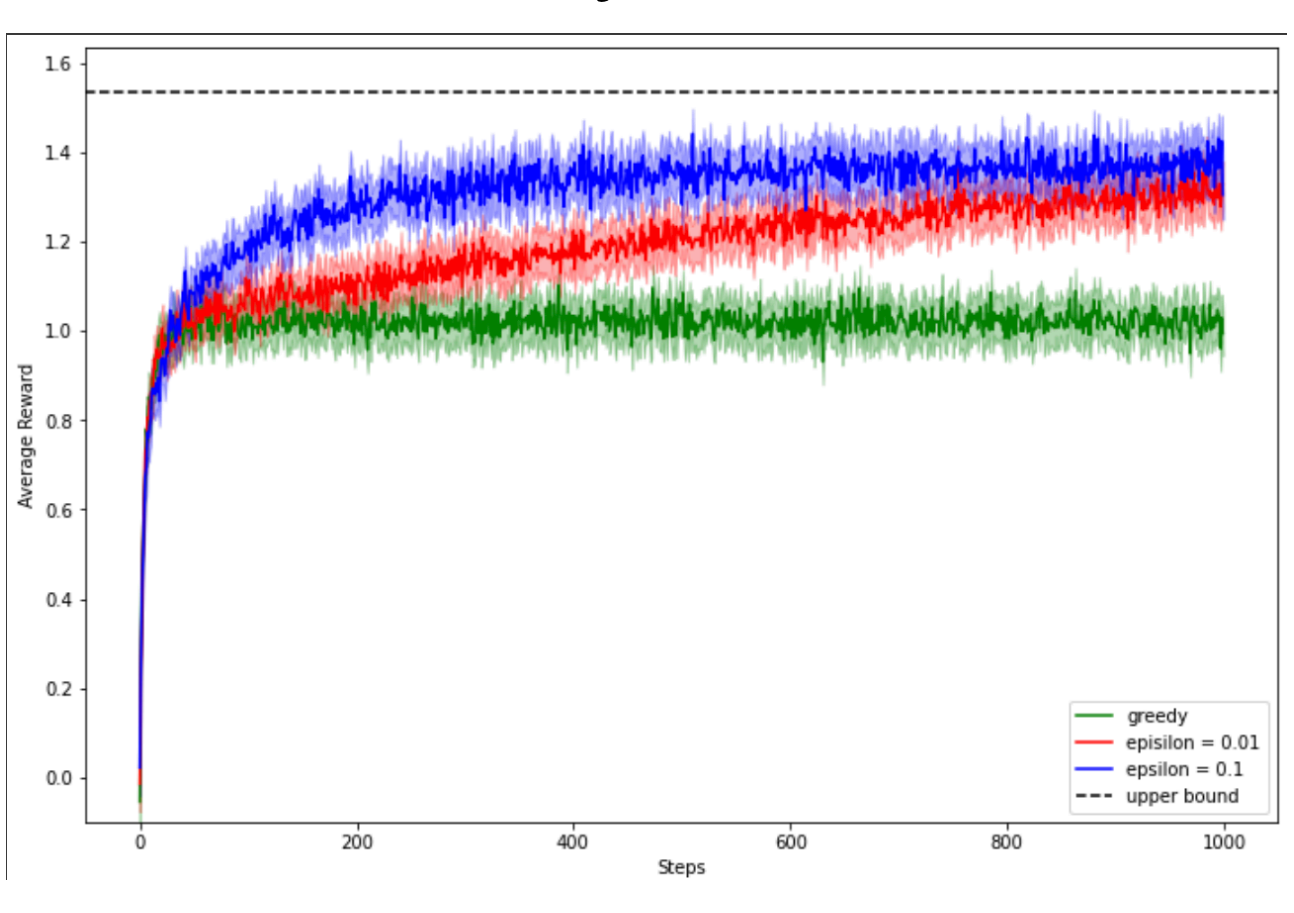
$$0.99 + 0.01 \times \frac{1}{10} = 99.1\%$$

Although greedy has 100% chances of Selecting best action due to lack of exploration it will get stuck

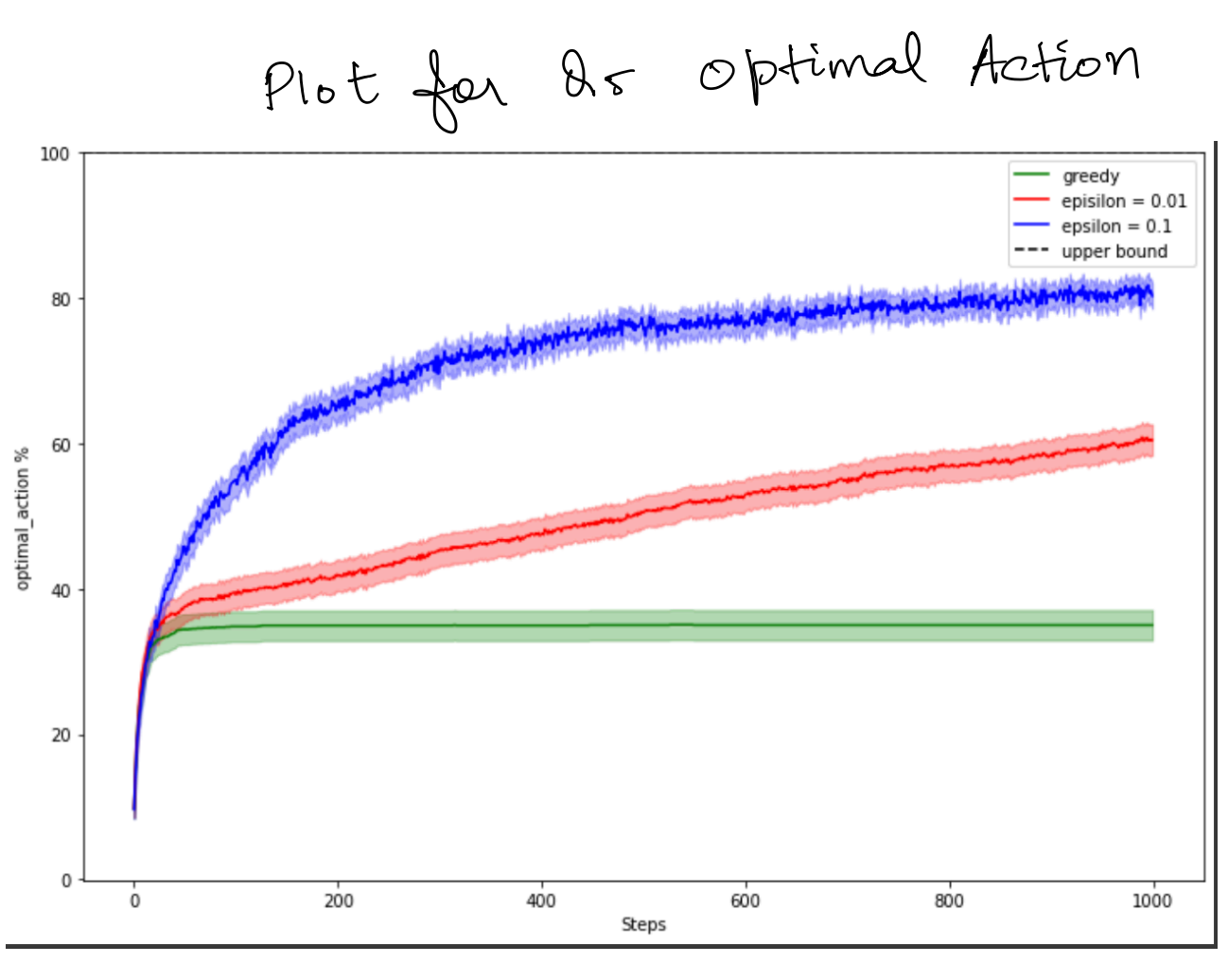
$\varepsilon = 0.1$ ← has more reward than $\varepsilon = 0.01$ initially but after some time when explored enough, It will also become More exploitative than exploratory

$\therefore$ $\varepsilon = 0.01$ Cumulatively will return higher reward

Plot Q5 for average Reward



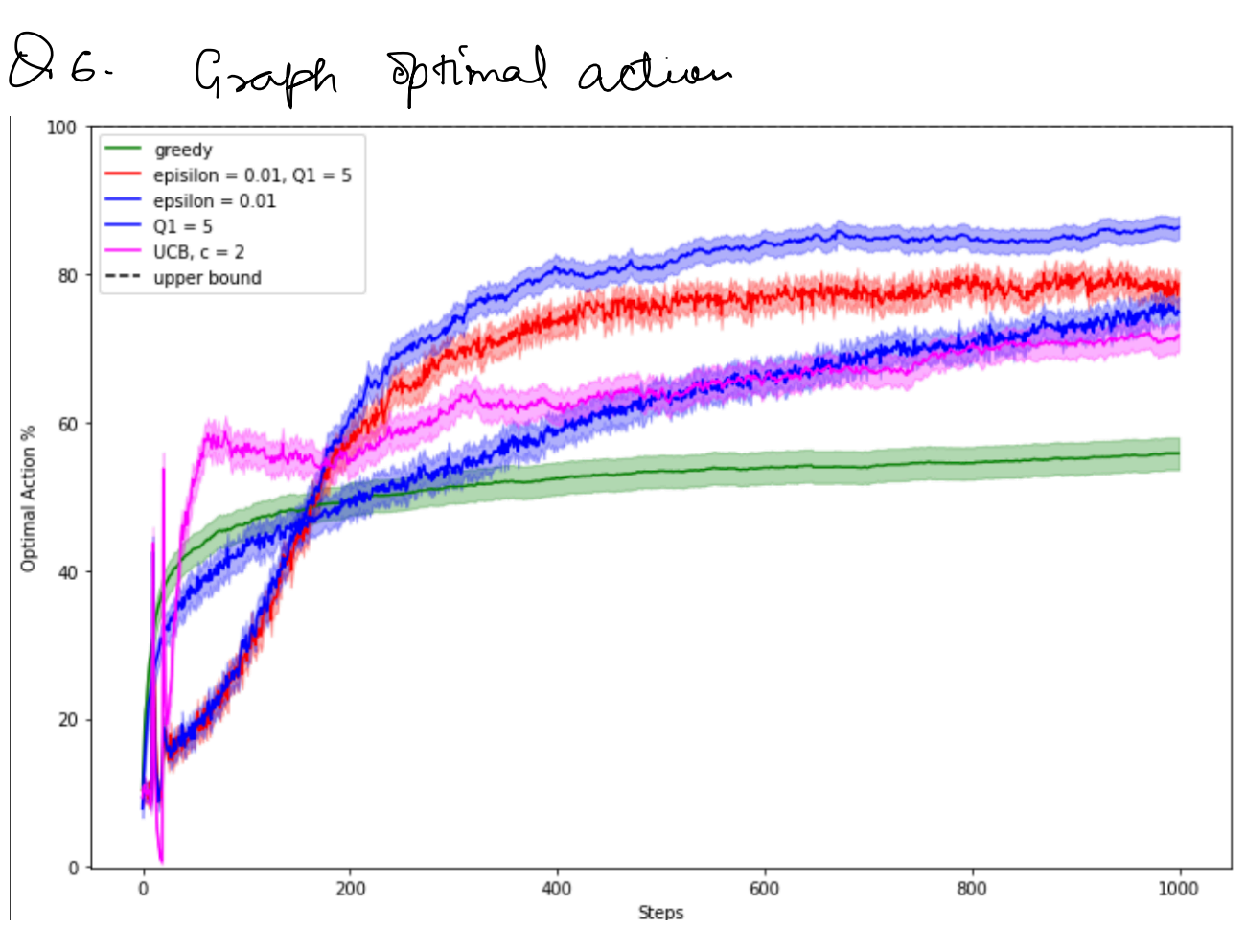
Plot for Q5 optimal Action



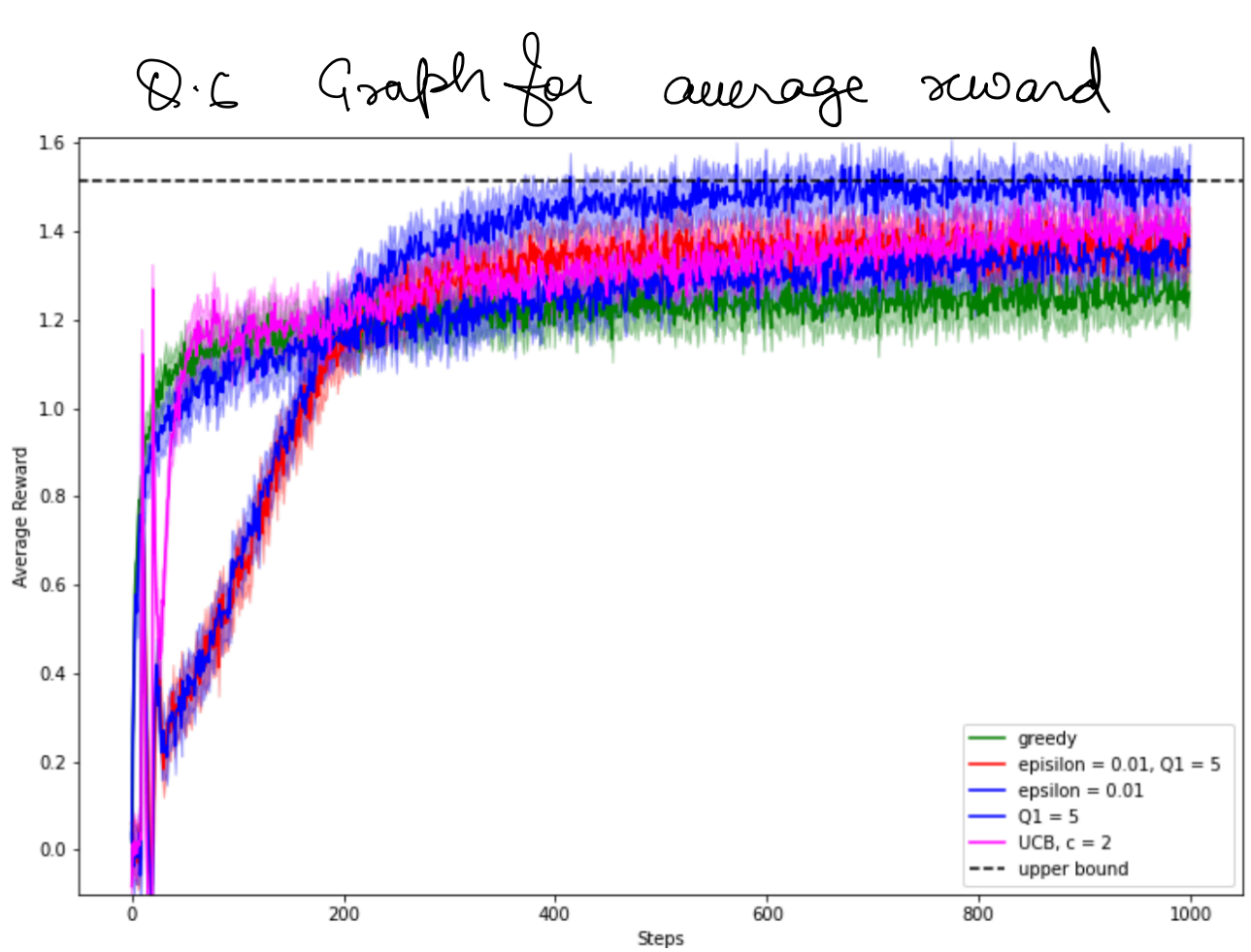
Q.6 written

we see the spike for the optimistic initialization in class.
Spike appears but not large enough
as $n \to \infty$, asymptomically it's possible

Q.6. Graph optimal action



Q.6 Graph for average reward



Curves with ε reach asymptotic & with initial Value has the spike