Garrett
Advanced Reinforcement Learning
February 1, 2022

Exercise 1

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 -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 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?

1) T1 $\rightarrow$ A1 = 1, R1 = -1 $\rightarrow$ A1 = -1, A2 = 0, A3 = 0, A4 = 0
2) T2 $\rightarrow$ A2 = 2, R2 = 1 $\rightarrow$ A1 = -1, A2 = 1, A3 = 0, A4 = 0
3) T3 $\rightarrow$ A3 = 2, R3 = -2 $\rightarrow$ A1 = -1, A2 = -1/2, A3 = 0, A4 = 0
4) T4 $\rightarrow$ A4 = 2, R4 = 2 $\rightarrow$ A1 = -1, A2 = 1/3, A3 = 0, A4 = 0 (definitely occurred)
5) T5 $\rightarrow$ A5 = 3, R5 = 0 $\rightarrow$ A1 = 1, A2 = 1/3, A3 = 0, A4 = 0 (definitely occurred)

Timesteps four and five definitely had the epsilon case occur since it chose an action with a sample-average action-value estimate lower than the maximum. In terms of time steps that epsilon possibly could have occurred, all time steps are possible, because random choice includes the maximum sample-average action-value choice, and thus even moves tzhat look optimal may have been random.

Question 2. If the step-size parameters, n, are not constant, then the estimate Qn 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?

$$
\begin{aligned}
Q_{n+1} &= Q_n + \alpha_n[R_n - Q_n] \\
&= (1-\alpha_n)Q_n + \alpha_n R_n \\
&\quad \textit{where } Q_n = (1-\alpha_{n-1})Q_{n-1} + \alpha_{n-1}R_{n-1} \\
Q_{n+1} &= (1-\alpha_n)(1-\alpha_{n-1})Q_{n-1} + (1-\alpha_n)\alpha_{n-1}R_{n-1} + \alpha_n R_n \\
&= Q_1 \prod_i^n (1-\alpha_i) + \sum_i^n \left(\alpha_i R_i \prod_{z=i}^{n-1} (1-\alpha_z) \right)
\end{aligned} \tag{1}
$$

Question 3. We say that an estimate is biased if the expected value of the estimate does not match the true value.

(a) Consider the sample-average estimate in Equation 2.1. Is it biased or unbiased? Explain briefly.
The bias caused by the the initial Q-value disappears once all actions have been selected at least once, while the bias of stochasticity converges to zero as time goes on and more samples are averaged into the overall estimate.

(b) If $Q_1 = 0$, is Q_n (for n > 1) biased? Explain briefly.
Unlike the previous answer, the bias of the initial value never disappears due to the constant weight applied to all samples, including the initial Q-value. However, in the case of $Q_1 = 0$, Q_n is canceled out in the (2.6) equation, so yes it is unbiased for Q_n.

(c) Derive condition(s) for Q_1 for when Q_n will be unbiased.
Q_n will be unbiased when Q_1 is cancels out in further downstream estimates of Q_n.

(d) Show that Q_n is an unbiased estimator as $n \rightarrow \infty$ (which is often referred to as asymptotically unbiased).
According to *pg. 33* of the textbook, the sum of the weights of the exponential recency-weighted average method equal 1. Thus the estimated for the expected reward value is a weighted average,

and by the law of large numbers in probability theory, this number converges to the true mean as the number of samples approaches infinity.

(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.

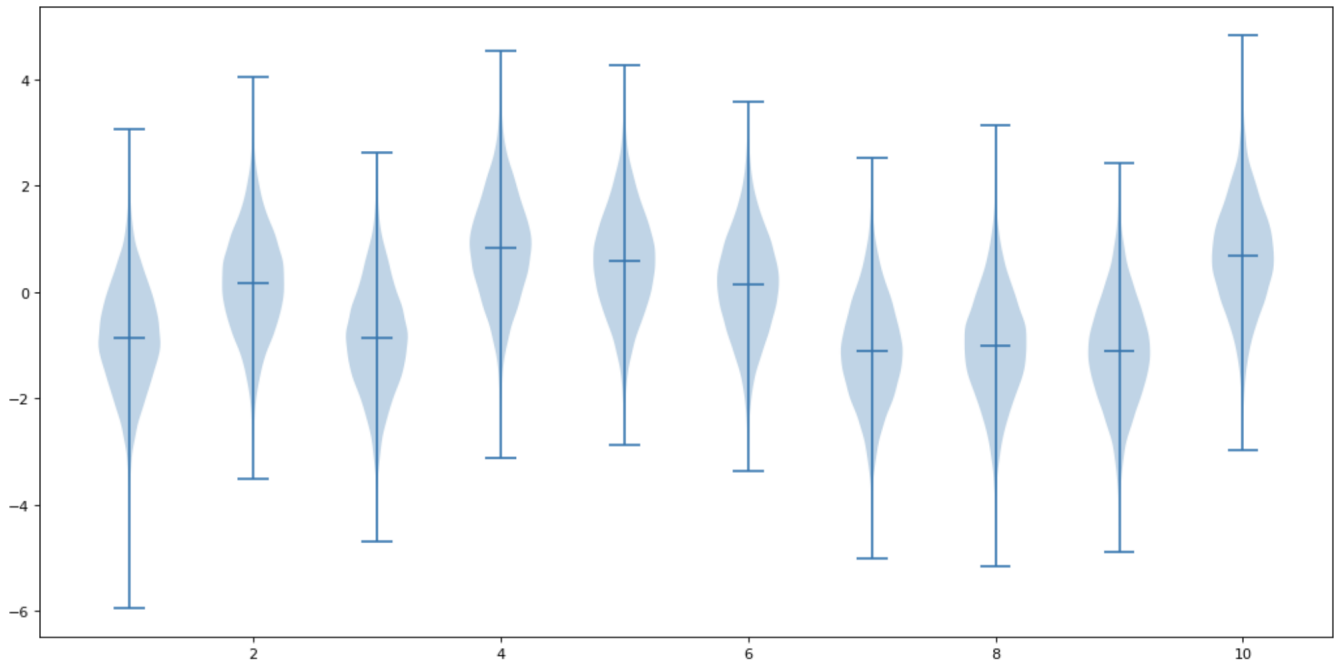
In practice, we will not have an infinite number of samples, thus the scenario described in *(d)* does not arise, and we have a situation in which $E[Q]! = q$, and the model is biased.

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.

Let $a = H_t(a_1)$ and $b = H_t(b_1)$.

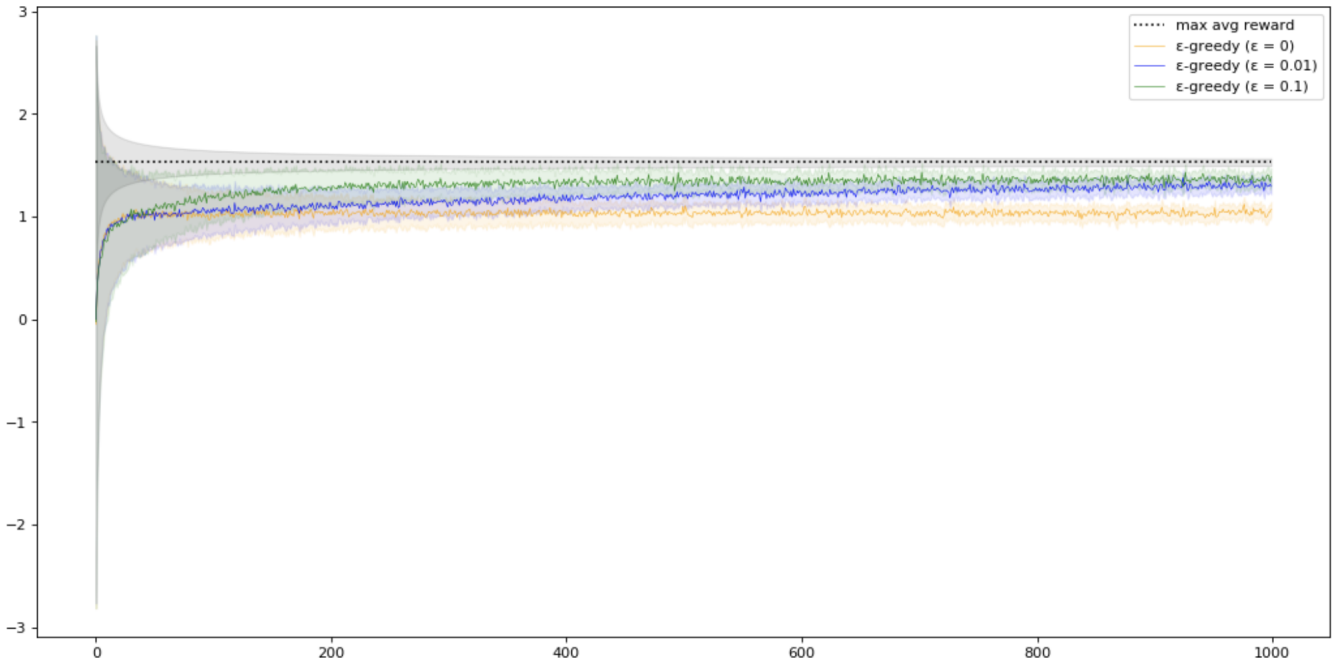
$$
\begin{aligned}
\text{softmax} &= \frac{e^a}{e^a + e^b} \\
&= \frac{1}{e^{-a}(e^a + e^b)} \\
&= \frac{1}{1 + e^{b-a}} \\
&= \frac{1}{1 + e^{b-a}} \\
&= \frac{1}{1 + e^{\theta}} \\
&= \sigma(\theta)
\end{aligned} \tag{2}
$$

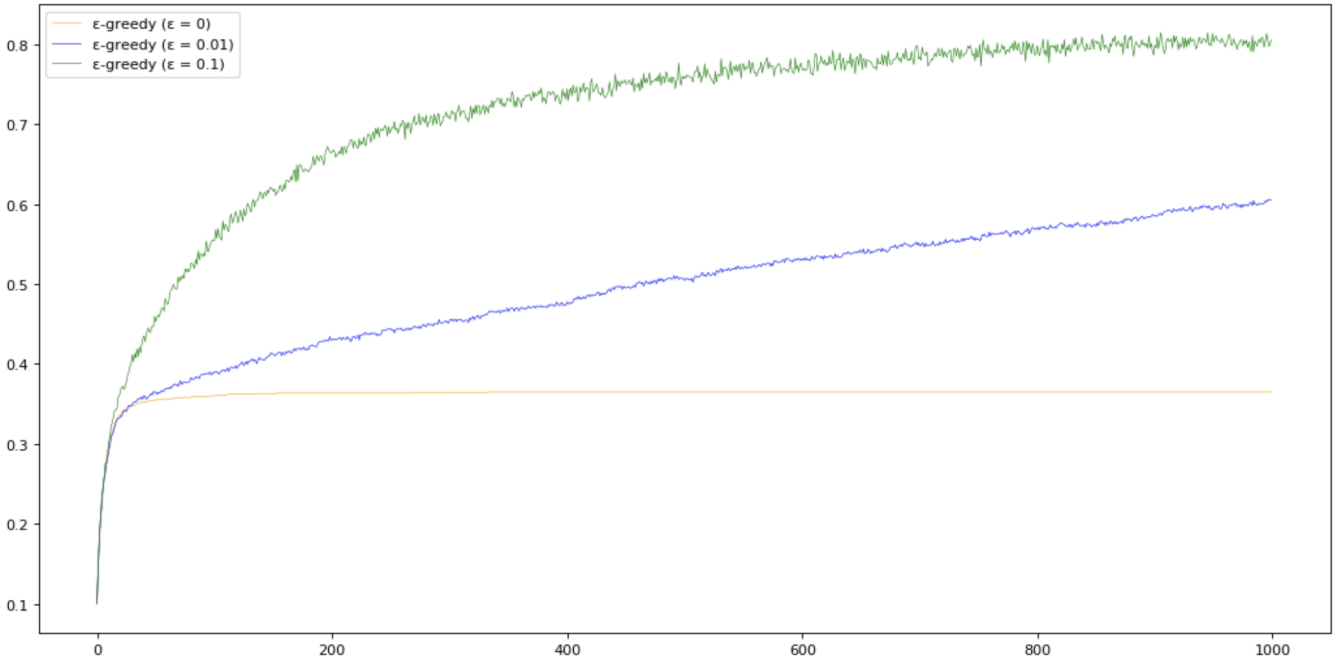
Question 5. To test that your testbed is working properly, produce a plot similar in style to Figure 2.1 by pulling each arm many times and plotting the distribution of sampled rewards. You can use any type of plot that makes this point effective, e.g., a violin plot, or a scatterplot with some jitter in the horizontal axis to show the sample density more effectively.



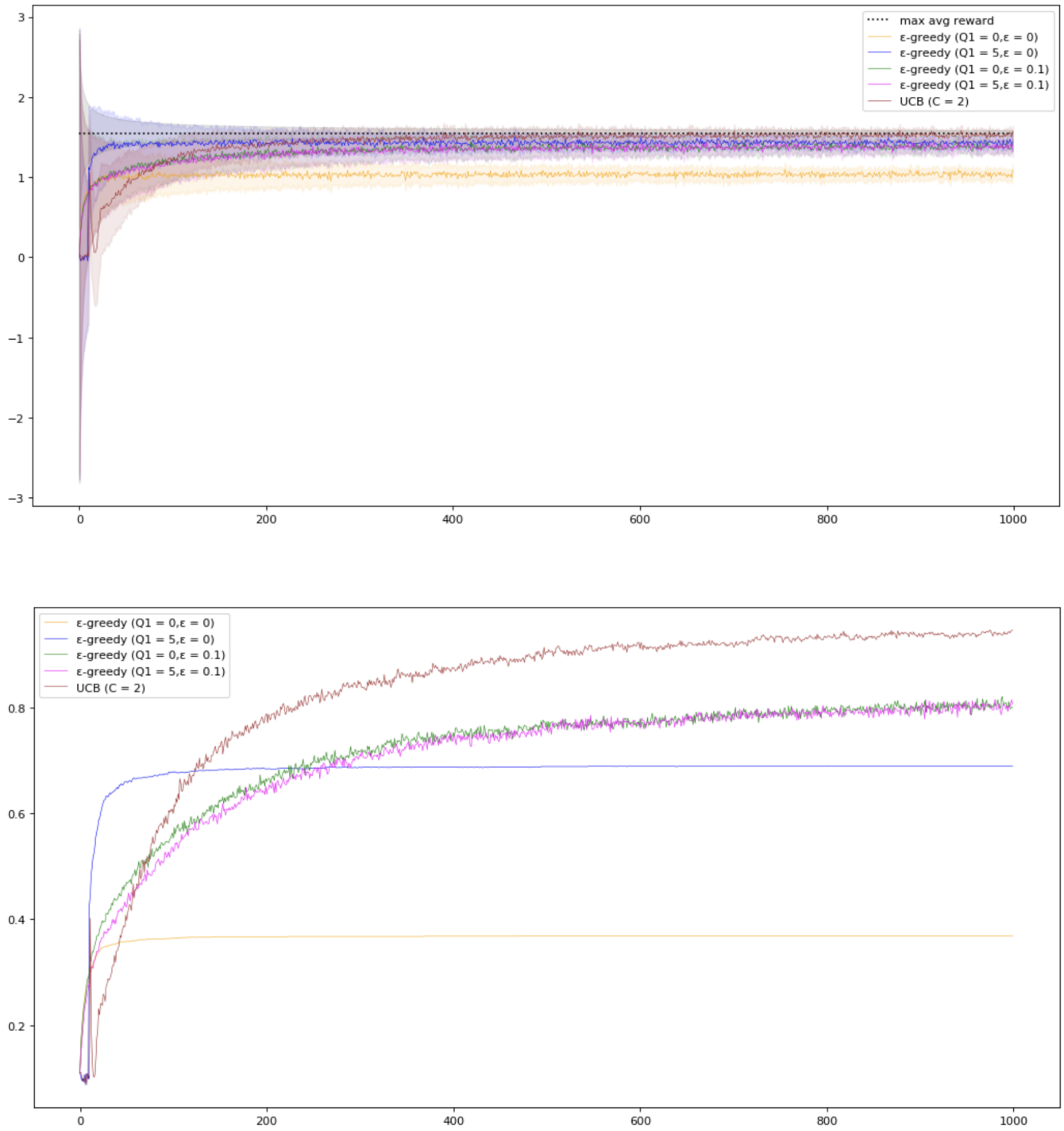
Question 6. Reproduce both plots shown in Figure 2.2 with the specified homework modifications. *Do the averages reach the asymptotic levels predicted in class?*

In terms of the plot that depicts the percentage that the model takes the optimal action, the answer is no. For $\epsilon = 0.01$ this would be 99% and for $\epsilon = 0.1$ this would be 90%. We need more runtime to reach these values.





Question 7. Reproduce the plots shown in both Figures 2.3 and 2.4 with the specified homework modifications.



We have seen the spike for optimistic initialization in class. Observe that UCB also produce spikes in the very beginning. 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.

UCB and optimistic ϵ-greedy algorithms produce sharp spikes in the beginning for the same reason, exploration, but as a consequence of different algorithm causes. For UCB, a denominator of zero causes all unexplored actions to appear maximal, and thus causes a high degree of exploration as apposed to exploitation in the beginning. Optimistic ϵ-greedy is similar, in that initializing high estimated Q values causes each actions estimated Q value to incrementally drop back down into the distribution range, thus doing a lot of exploring in the initial steps. We can empirically back this up by tracking how many times

the method switches its view on the most optimal action to take (exploration), and see that in initial stages UCB and optimistic ϵ-greedy have high exploration in beginning steps.