Luke Davidson
CS 5180
Ex. 1

1) $Q_t(a)$

t	1	2	3	4	A_t	R_t
1	0	0	0	0	1	-1
2	-1	0	0	0	2	1
3	-1	1	0	0	2	-2
4	-1	-0.5	0	0	2	2
5	-1	0.33	0	0	3	0

$Q_4(2) = \frac{1-2}{2} = -\frac{1}{2}$ $\quad Q_5(2) = \frac{1-2+2}{3} = \frac{1}{3} = 0.33$

$t=1$: May have occured. Since all $Q_1(a)$ are 0, both the ε and $1-\varepsilon$ options would select arbitrarily, so we don't know which it was.

$t=2$: May have occured. Similar to $t=1$, $Q_2(\{2,3,4\})$ are all the max and $=0$. For ε case, a random selection of 1, 2, 3 or 4 would be made, so 2 would be possible. For $1-\varepsilon$ case, a random selection of 2, 3 or 4 would be made, so 2 also possible.

$t=3$: May have occured. For ε case, a random selection of 1-4 would have been made, so 2 possible. For $1-\varepsilon$ case, 2 would be selected since it is the max $Q_3(a)$.

$t=4$: Definitely occurred. $1-\varepsilon$ case would have selected 3 or 4 since the $Q_4(\{3,4\})$ are maxes. Since 2 was selected, we know it was by random, or ε case.

$t=5$: Definitely occurred. $1-\varepsilon$ case would have selected 2 since 0.33 is max $Q_5(a)$. 3 was selected so we know it was by random, or ε case.

2) $Q_{n+1} = Q_n + \alpha_n [R_n - Q_n]$

$= Q_n + \alpha_n R_n - \alpha_n Q_n$

$= \alpha_n R_n + Q_n (1 - \alpha_n)$

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

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

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

$= (\,) \cdot (Q_{n-2} + \alpha_{n-2}[R_{n-2} - Q_{n-2}])$

$= (\,) \cdot R_{n-2}\alpha_{n-2} + (1-\alpha_{n-2})(Q_{n-2})$

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

eventually will lead to

$$Q_{n+1} = \left[\prod_{i=1}^{n} (1-\alpha_i) \right] Q_1 + \sum_{i=1}^{n-1} \alpha_i \prod_{i-1}^{n} (1-\alpha_{i-1}) R_i + \alpha_n R_n$$

3) a) eq. 2.1 = $\frac{\sum_{i=1}^{t-1} R_i \, \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$ =

$E[Q_n]$

$Q_{t+} = E[R_t \mid A_t = a]$

$E[Q_n] =$

$\hookrightarrow Q_n = \frac{R_1 + \ldots + R_{n-1}}{n-1}$ $\Longrightarrow E\left[\frac{R_1 + \ldots + R_{n-1}}{n-1}\right]$

$\longrightarrow = E\left[\frac{\sum_{i=1}^{n-1} R_i}{\sum_{i=1}^{n-1} \mathbb{1}_{A_i=a}}\right]$

for eq. 2.1

$E\left[\frac{\sum_{i=1}^{t-1} R_i \, \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}\right]$

Comparing both, they are equivalent through equating n and t. so it is unbiased.

3b) eq 2.5) $Q_{n+1} = Q_n + \alpha[R_n - Q_n]$

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

$q_* = E[R_t \mid A_t = a]$

$= \sum p(x) \cdot x$

probability of selection

for $Q_1 = 0$

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

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

$E[Q_n] = E\left[\sum_{i=1}^{n-1} \alpha(1-\alpha)^{n-1-i} R_i\right]$

$= E\left[\frac{\sum R}{n-1}\right] \neq \sum_{i=1}^{n-1} E[R_i]\,\alpha(1-\alpha)^{n-1-i}$ } this statement is not true, so when $Q_1 = 0$, the exponential recency-weighted average estimate is biased.

3c.) for Q_n to be unbiased

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

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

must solve for a Q_1, or rule for Q_1, where

$E[Q_n] = E\left[\frac{\sum_{i=1}^{n-1} R_i}{n-1}\right] = q_* = E[R_t \mid A_t = a]$

$$Q_1 = \frac{E\left[\frac{\sum_{i=1}^{n-1} R_i}{n-1}\right] - \sum_{i=1}^{n-1} E[R_i]\,\alpha(1-\alpha)^{n-i}}{E[(1-\alpha)^{n-1}]}$$

$n \to \infty$

3d.) $Q_{n+1} = Q_n + \alpha[R_n - Q_n]$

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

as $n \to \infty$,

$(1-\alpha)^{\infty} Q_1 + \sum_{i=1}^{\infty} \alpha(1-\alpha)^{\infty - i} R_i \to$ (0) b/c $0 < \alpha < 1$, so $(1-\alpha)^{\infty}$ will continue to get smaller and smaller and eventually get to 0.

$E[Q_n] = E\left[\frac{\sum_{i=1}^{n-1} R_i}{n-1}\right]$ as $n \to \infty$

also $\to$ (∞) b/c the denom will be ∞. Therefore as $n \to \infty$ Q_n is <u>unbiased</u> since both = 0.

3e). We should expect the exponential recency-weighted average will be biased in practice due to the original effect Q_1 has on it. For

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

as n increases, $(1-\alpha)^n$ will become smaller and smaller (given $0 < \alpha < 1$), although never 0. Because it will never achieve the value of 0, there will always be a diminishing bias towards Q_1, thus it will behave as a bias estimator.

4.) softmax $\Pr\{A_t = a\} = \dfrac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} = \pi_t(a)$

logistic $\dfrac{L}{1 + e^{-k(x - x_0)}}$

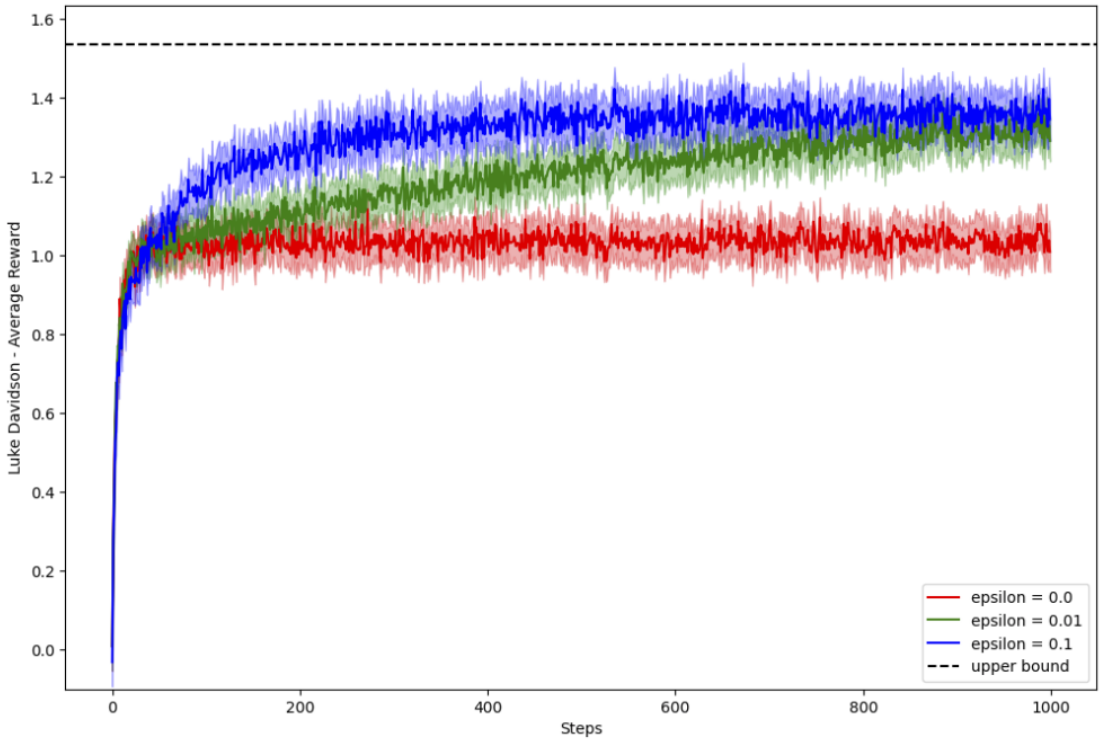
2 actions $\Pr\{A_t = a\} = \dfrac{e^{H_t(a)}}{e^{H_t(a)} + e^{H_t(b)}} \cdot \dfrac{\frac{1}{e^{H_t(a)}}}{\frac{1}{e^{H_t(a)}}}$

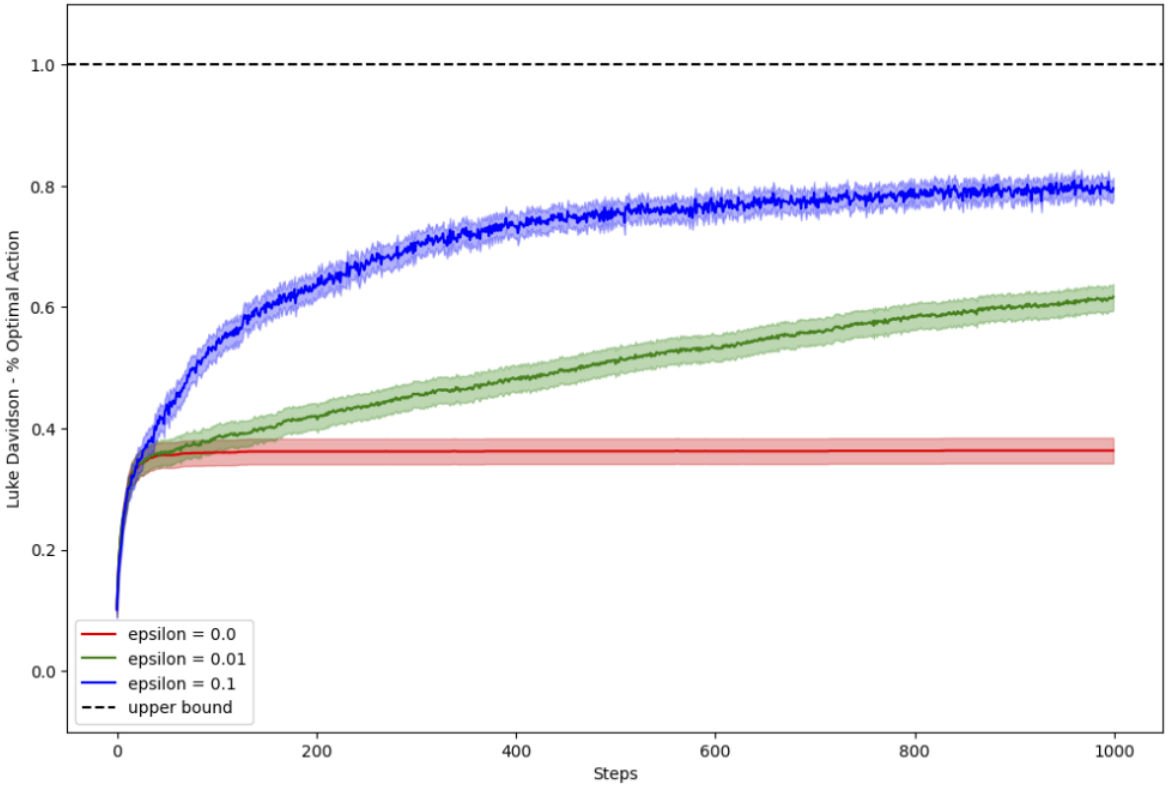
$\Rightarrow \boxed{\dfrac{1}{1 + e^{H_t(b) - H_t(a)}}}$ which is in same form as

logistic $\boxed{\dfrac{L}{1 + e^{-k(x - x_0)}}}$ where $L = 1$ and $-k(x - x_0) = H_t(b) - H_t(a)$

✓

5 Plot

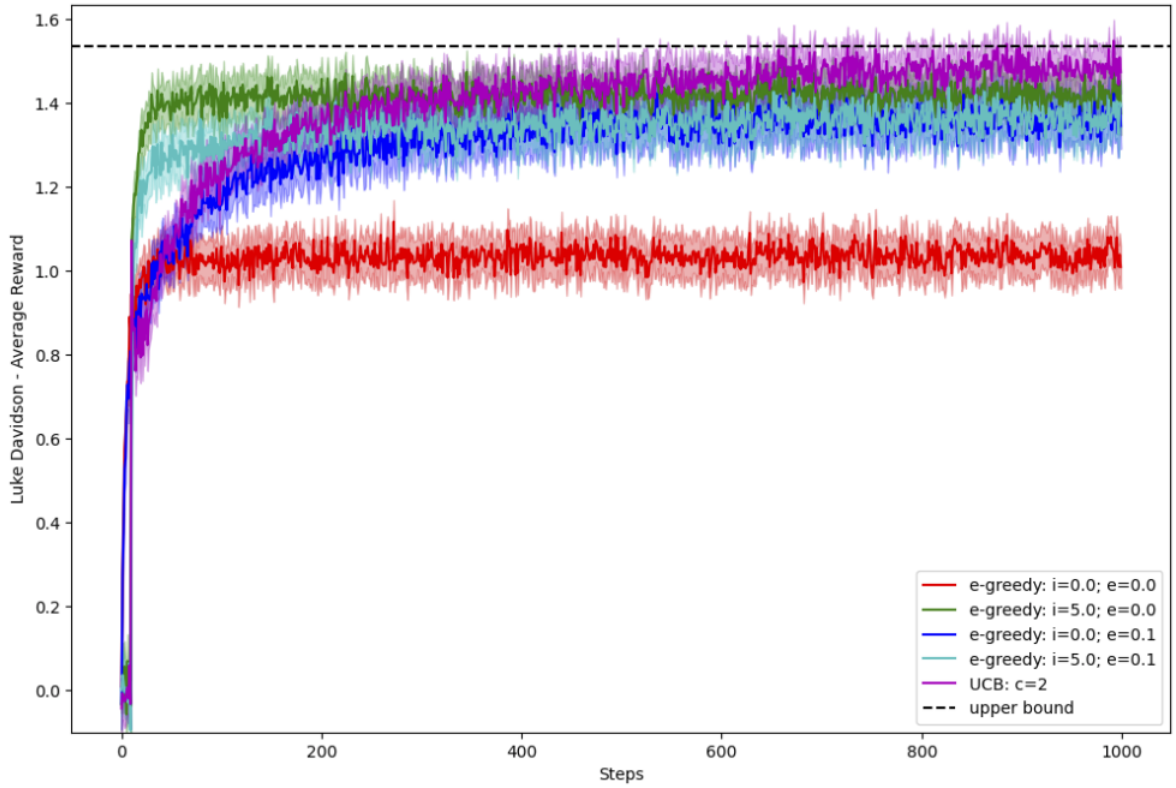

Luke Davidson - Average Reward
Steps
epsilon = 0.0
epsilon = 0.01
epsilon = 0.1
upper bound

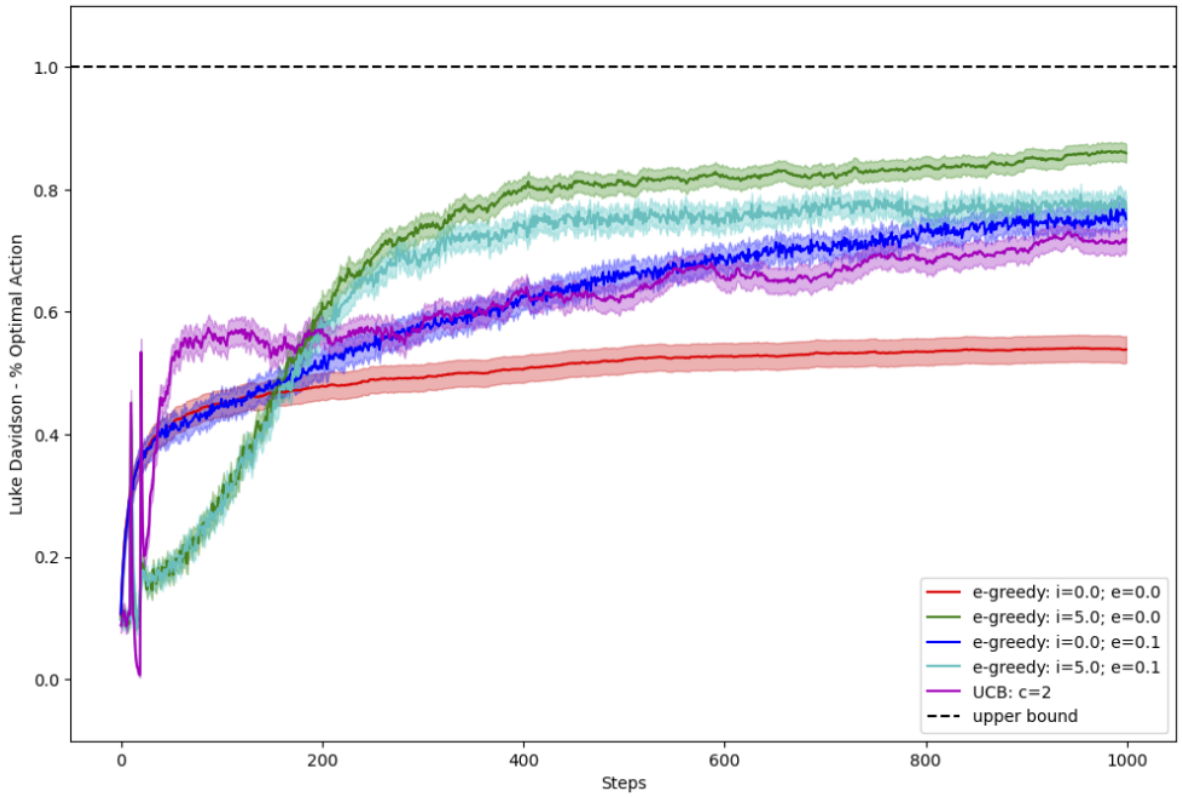

Luke Davidson - % Optimal Action
Steps
epsilon = 0.0
epsilon = 0.01
epsilon = 0.1
upper bound

5.) Written: The average rewards that the plots converge to essentially represent the upper limit each of those individual policies would be able to reach (on average). As the # of steps increase, eventually to infinity, each of these policies will have different levels of optimality, represented by the values they converge to. At some point, each policy will not be able to gain more rewards per action.

This is due to the nature of the policies, which itself is due to the hyperparameters. These policies behave differently with different epsilon values. A policy with a high epsilon value will more often times land on a random action choice compared to lower epsilon values, thus "exploring" more and "exploiting" less. A perfect ratio of exploring and exploiting is desired to obtain a higher optimal convergence. Too much exploration and the agent will not exploit its knowledge of which action may be best enough. Too much exploitation and the agent will not learn the environment well enough. This is seen in our plots with $\varepsilon = \{0, 0.01, 0.1\}$.

6 Plot


Luke Davidson - Average Reward
Steps
e-greedy: i=0.0; e=0.0
e-greedy: i=5.0; e=0.0
e-greedy: i=0.0; e=0.1
e-greedy: i=5.0; e=0.1
UCB: c=2
upper bound


Luke Davidson - % Optimal Action
Steps
e-greedy: i=0.0; e=0.0
e-greedy: i=5.0; e=0.0
e-greedy: i=0.0; e=0.1
e-greedy: i=5.0; e=0.1
UCB: c=2
upper bound

(c.) Written: At the beginning rewards are low because all actions are treated as maxes due to $N_t(a) = 0$, so they are selected randomly. With a higher "c", the variance in the estimate of a's value is amplified. After all the actions are selected an initial time, eventually a high reward action will be selected randomly, which will greatly increase the average reward of that step, resulting in the spike seen. After the first spike, and similarly to the first part, "c" acts to not allow the agent to act in a greedy way, so it will likely not select the highest reward action again. This results in the decrease after the initial spike. As t grows, this will have less of an affect on the results. Similarly as Q continues to be updated.

This pr... ress more or less so continues until...