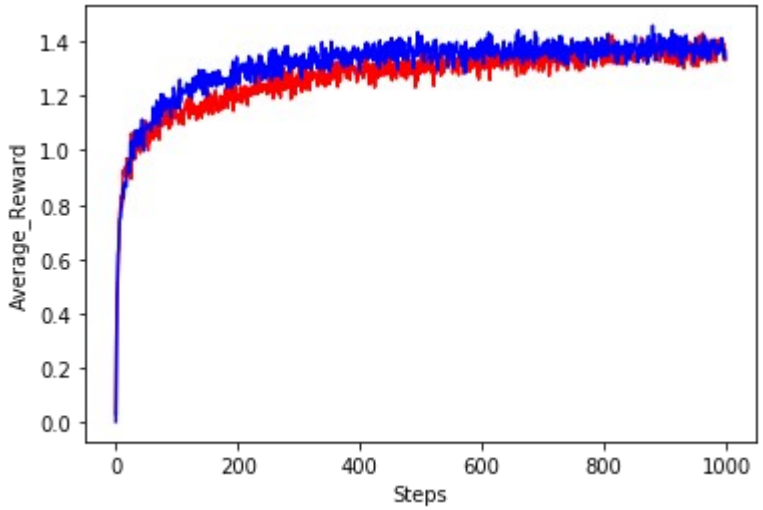
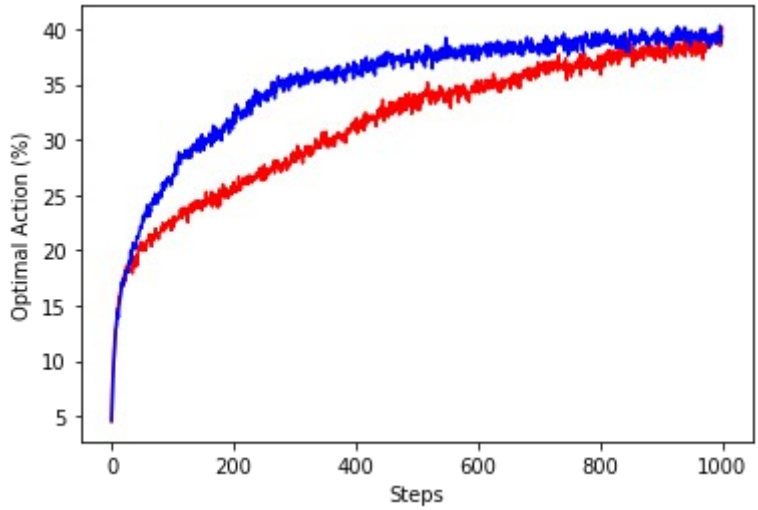
Assignment 1

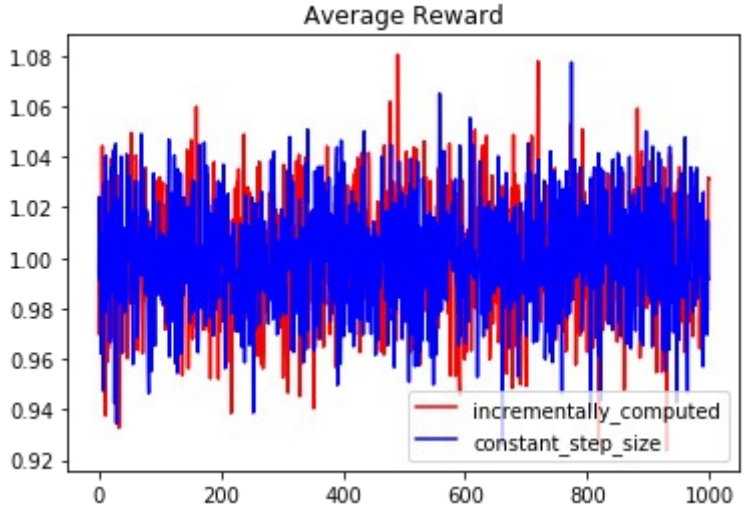
Question 1

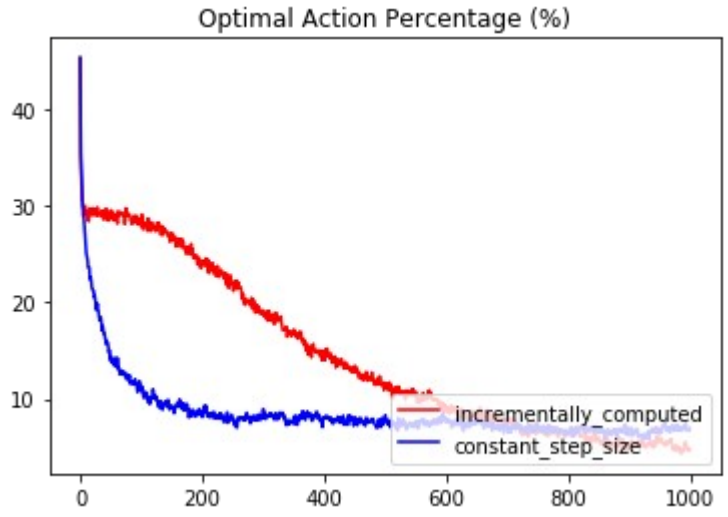
q = np.random.normal(0,1,n_arms)





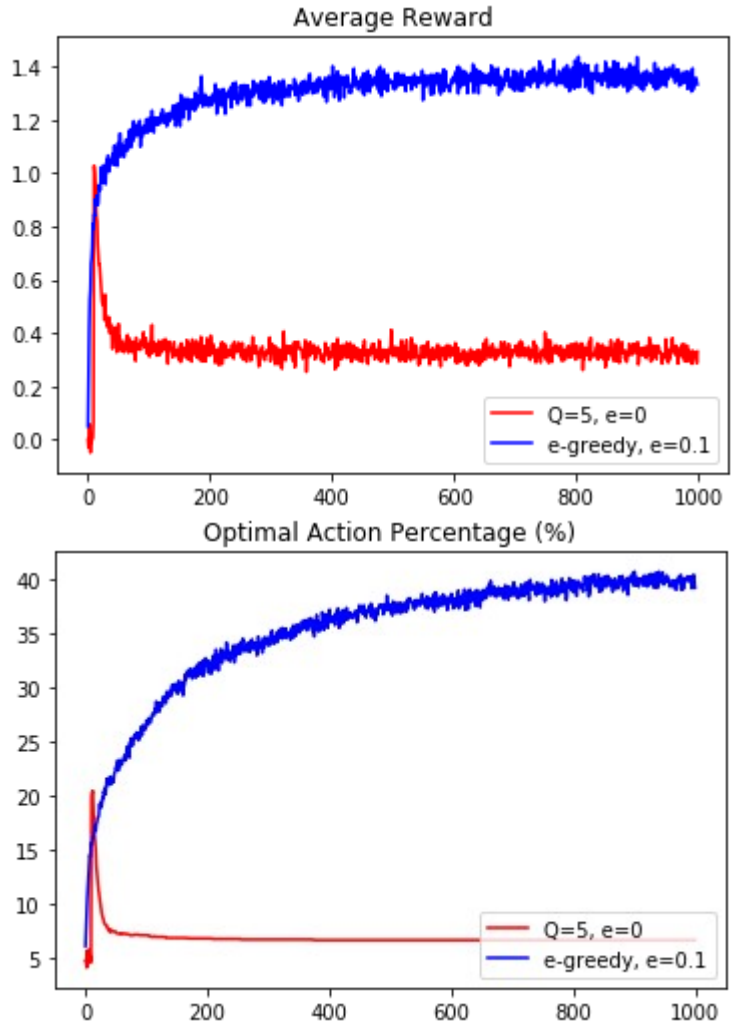
q = np.ones(n_arms)



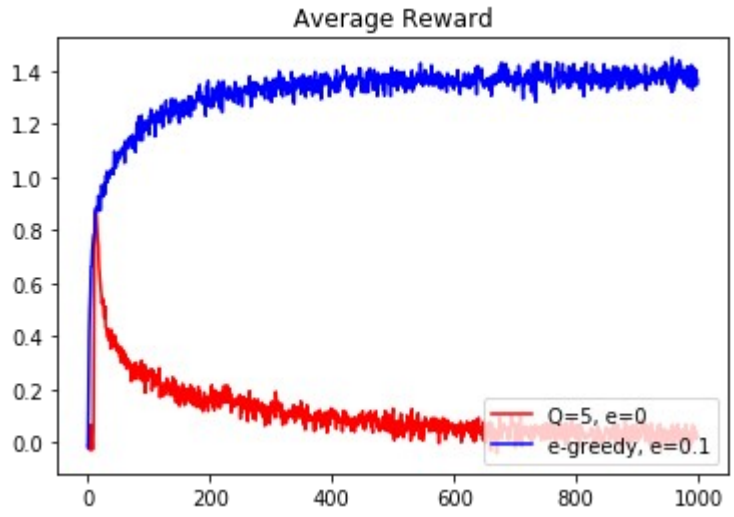

Optimal Action Percentage (%)
40
30
20
10
0
200
400
600
800
1000
incrementally_computed
constant_step_size

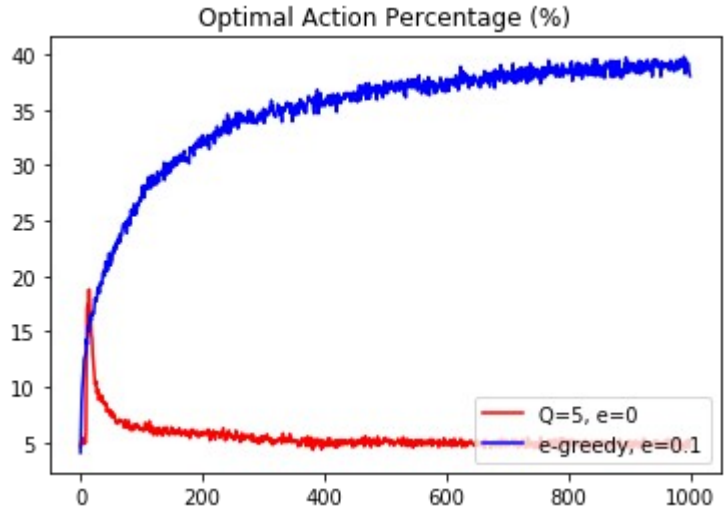
Question 2

Question 2 fig 2.3 for stationary ; incrementally computed

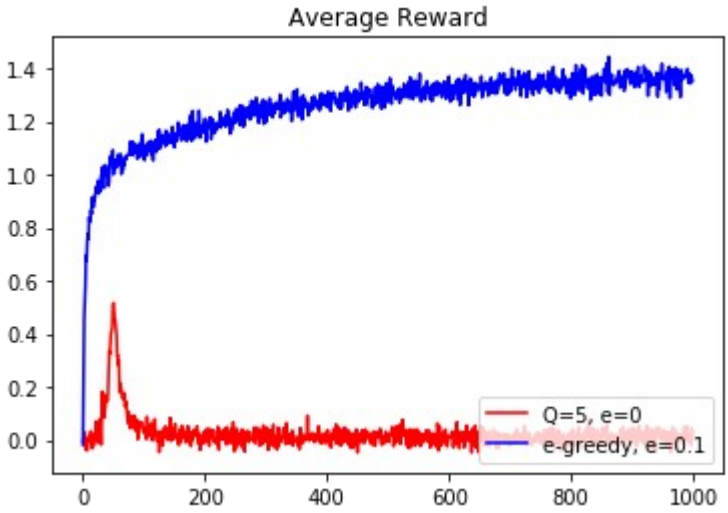


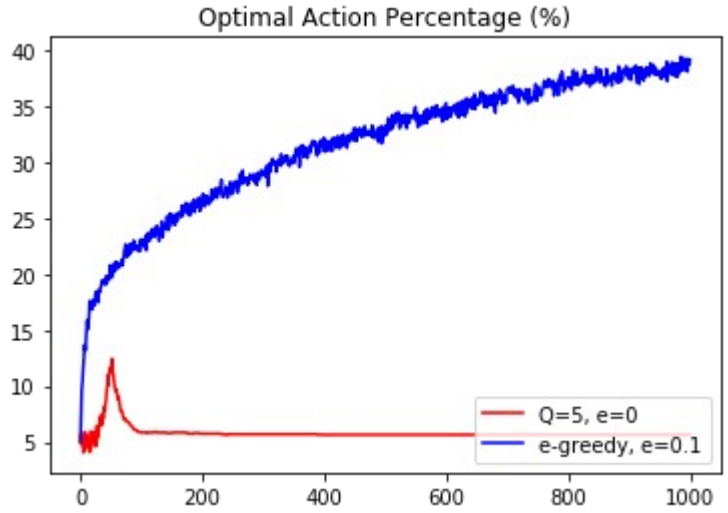
Question 2 fig 2.3 for ; non-stationary; incrementally-computed



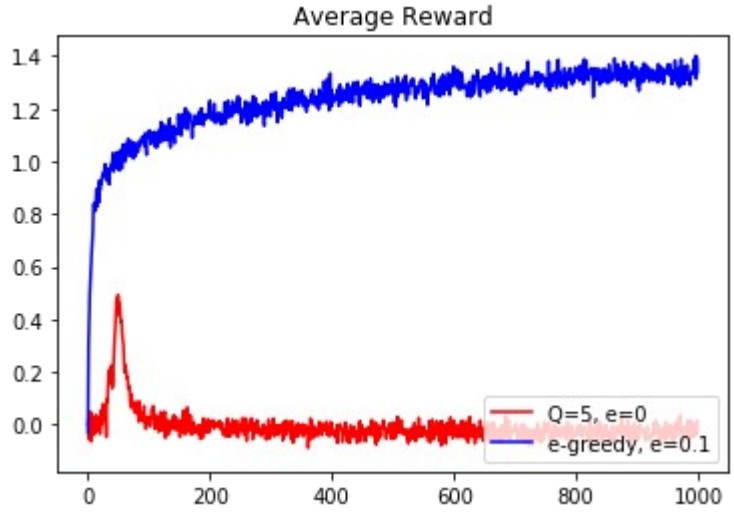


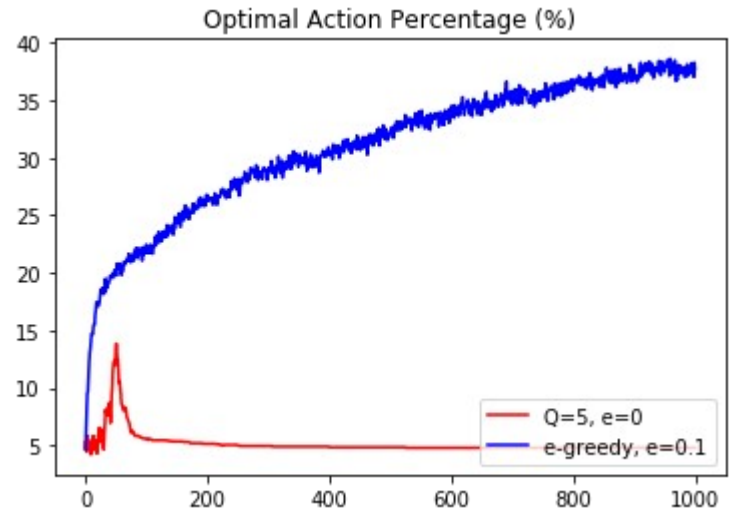
Question 2 fig 2.3 for stationary ; constant_step-size





Question 2 fig 2.3 for non stationary ; constant_step-size





Initially, the optimistic method performs worse because it explores more,but eventually it performs better because its exploration decreases with time.

Solution for question 6

The oscillations in start of optimistic curve show that estimated values are still optimistic. When the curve increases, it is due to choosing the the bandit with the highest expected reward. Estimated values are still optimistic (much higher than their expected values). In this context, the best bandit is selected due to its optimistic value during the early part of the curve, but when its estimated value becomes much closer to the real expected value, other decisions with optimistic values are chosen.

Due to this the curve oscillates during the early stages, until all estimated values aren't as nearly as optimal and become much closer to the real expected value instead.

Thus optimistic greedy is not better for small number of time steps. Realistic e-greedy performs better for that. Optimistic method encourage exploration only during the beginning of execution, and would just be useful for stationary problems.

Question 3:

$$Q_{n+1} = Q_n + \beta_n [R_n - Q_n]$$

$$Q_{n+1} = Q_n + \frac{\alpha}{\bar{o}_n} [R_n - Q_n]$$

$$= \frac{\alpha}{\bar{o}_n} R_n + \left(1 - \frac{\alpha}{\bar{o}_n}\right) Q_n$$

$$= \frac{\alpha}{\bar{o}_n} R_n + \left(1 - \frac{\alpha}{\bar{o}_n}\right) \left[\frac{\alpha}{\bar{o}_{n-1}} R_{n-1} + \left(1 - \frac{\alpha}{\bar{o}_{n-1}}\right) Q_{n-1}\right]$$

$$= \frac{\alpha}{\bar{o}_n} R_n + \left(1 - \frac{\alpha}{\bar{o}_n}\right)\left(\frac{\alpha}{\bar{o}_n}\right) R_{n-1} + \left(1 - \frac{\alpha}{\bar{o}_n}\right)^2 Q_{n-1}$$

$$= \frac{\alpha}{\bar{o}_n} R_n + \left(1 - \frac{\alpha}{\bar{o}_n}\right)\left(\frac{\alpha}{\bar{o}_n}\right) R_{n-1} + \left(1 - \frac{\alpha}{\bar{o}_n}\right)^2 \frac{\alpha}{\bar{o}_n} R_{n-2} +$$

$$\cdots + \left(1 - \frac{\alpha}{\bar{o}_n}\right)^{n-1} \cdot \frac{\alpha}{\bar{o}_n} R_1 + \left(1 - \frac{\alpha}{\bar{o}_n}\right)^n Q_1$$

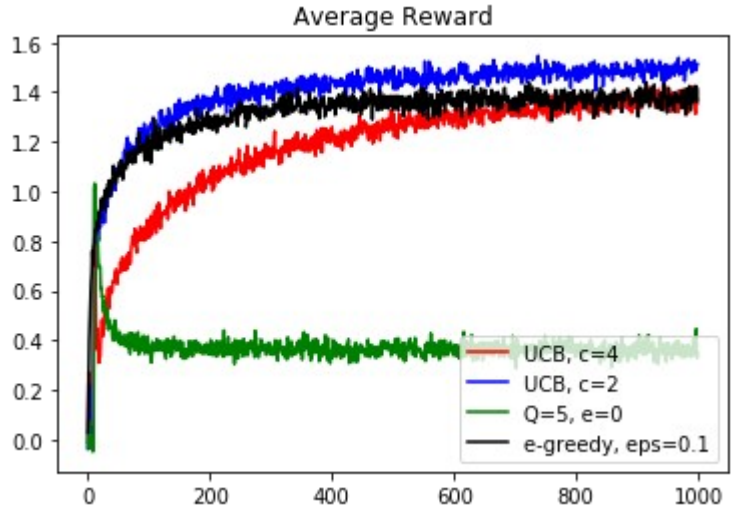
$$= \left(1 - \frac{\alpha}{\bar{o}_n}\right)^n Q_1 + \sum_{i=1}^{n} \frac{\alpha}{\bar{o}_n} \left(1 - \frac{\alpha}{\bar{o}_n}\right)^{n-i} R_i$$

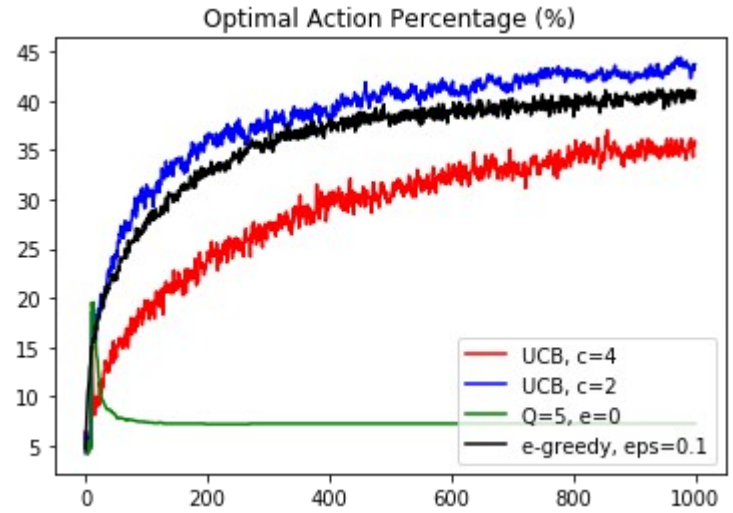
$$= \frac{1}{\left(\frac{\bar{o}_n}{\alpha}\right)} \left[\underbrace{(1-\alpha)^n \bar{o}_0 Q_1}_{=0} + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i\right]$$

Since $\bar{o}_0 = 0$ hence bias $= 0$ Thus $Q_{n+1} = \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i$

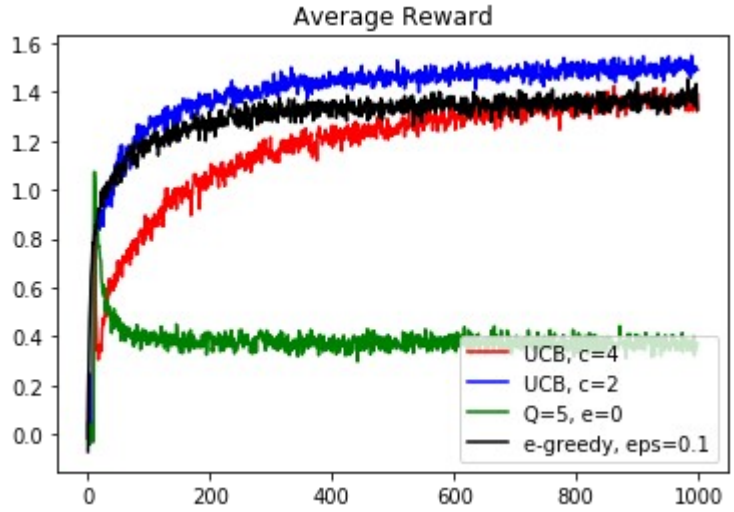
Question 4

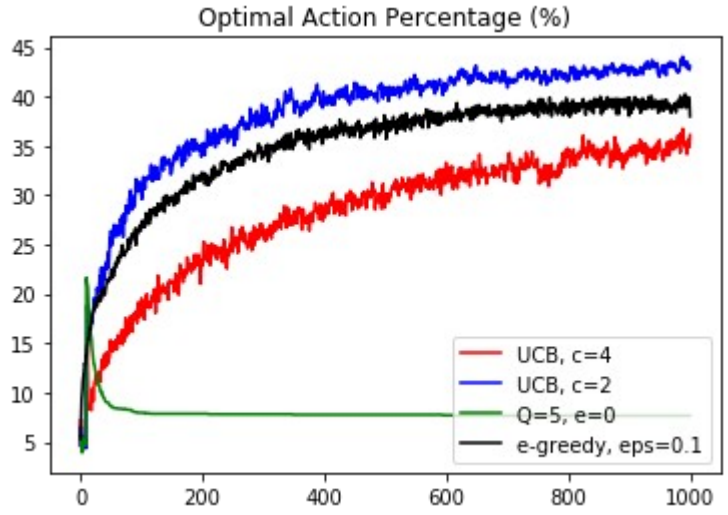
UCB for c=4, c=2, e-greedy = 0.1, Q=5, e=0 ; incrementally computed ; stationary



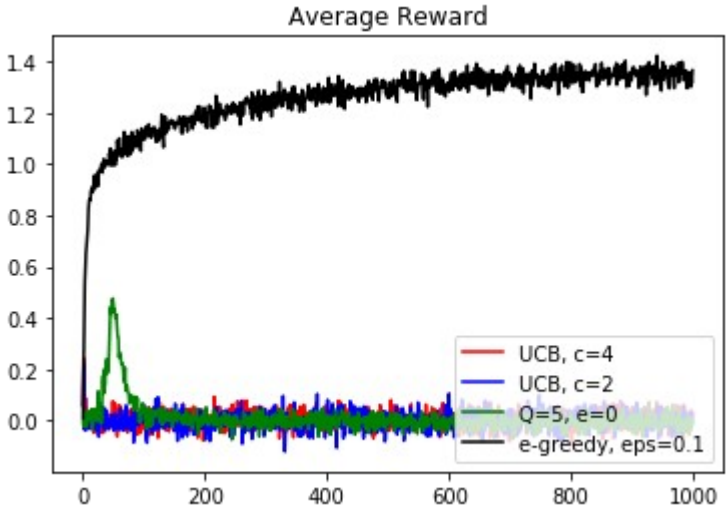


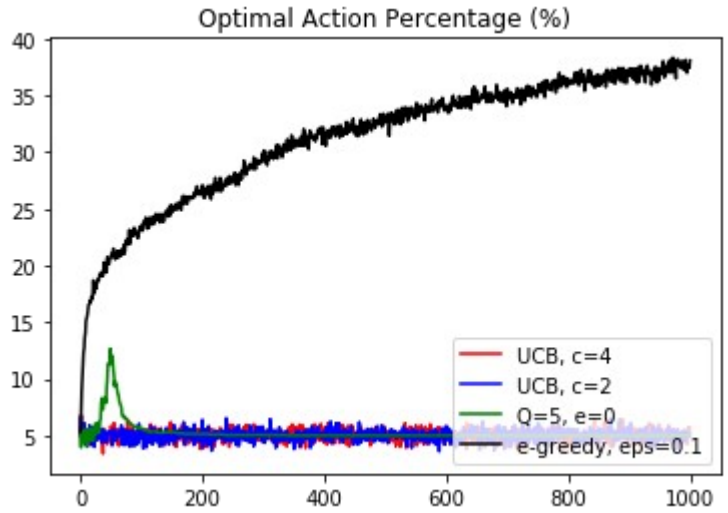
UCB for c=4, c=2, e-greedy = 0.1, Q=5, e=0 ; incrementally computed ; non-stationary



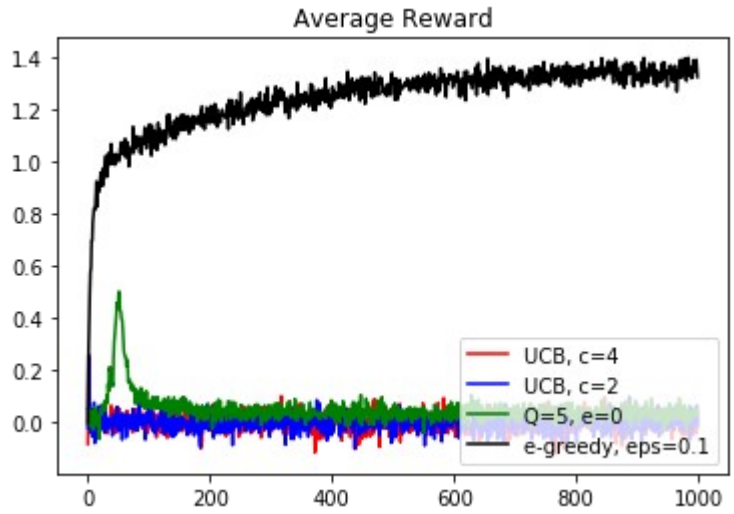


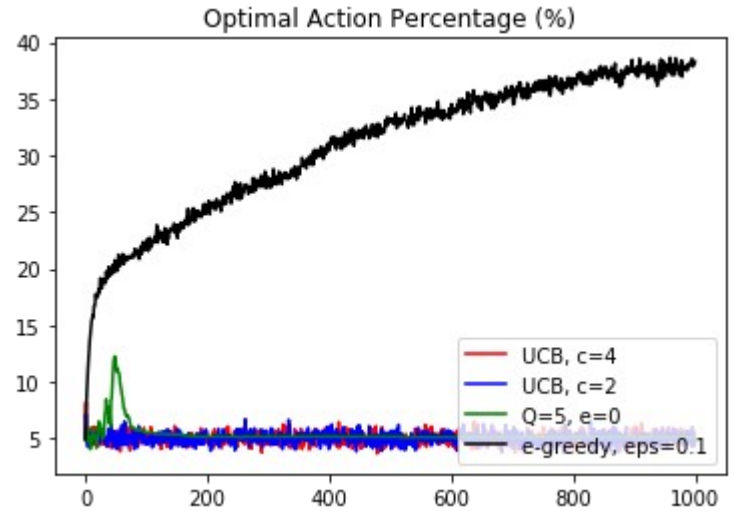
UCB for c=4, c=2, e-greedy = 0.1, Q=5, e=0 ; constant_step-size ; stationary





UCB for c=4, c=2, e-greedy = 0.1, Q=5, e=0 ; constant_step-size; non-stationary





UCB performs better. For lower value of c its performance is better. Its performance is poor for constant step size.
E-greedy performs best for constant step size.
Optimistic greedy has a initial increased spike but after that its performance decreases.