REINFORCEMENT LEARNING - CSE564

Homework 1

Suchet Aggarwal

2018105

Q1. The Epsilon Value chosen for the converging case is min(50/t,1) where t = 1,2,3,4…., this choice of epsilon forces the agent to explore its surroundings with probability 1 for atleast the first 50 time steps, this helps in estimating the estimate reward well for all arms atleast for 50 time steps, after which the epsilon decreases and follows equation 2.7.

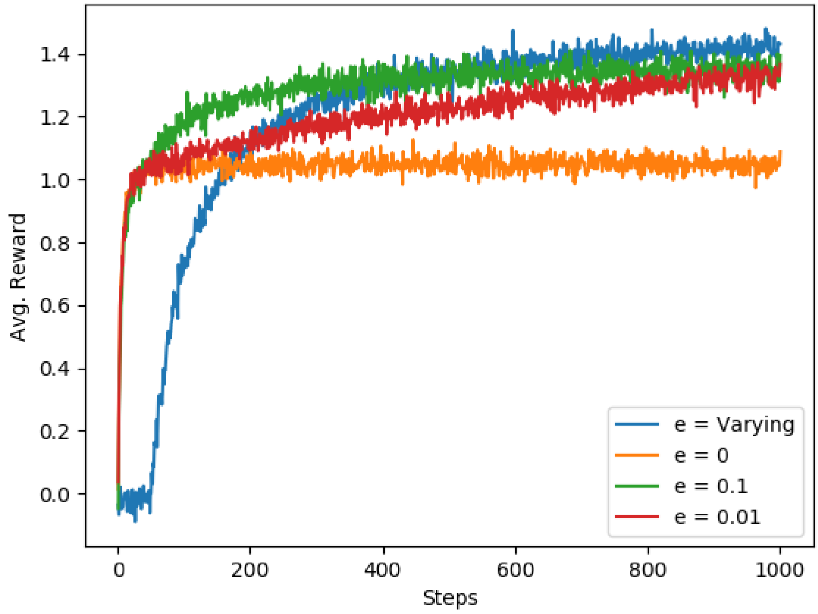

Fig 1.1: Average Reward

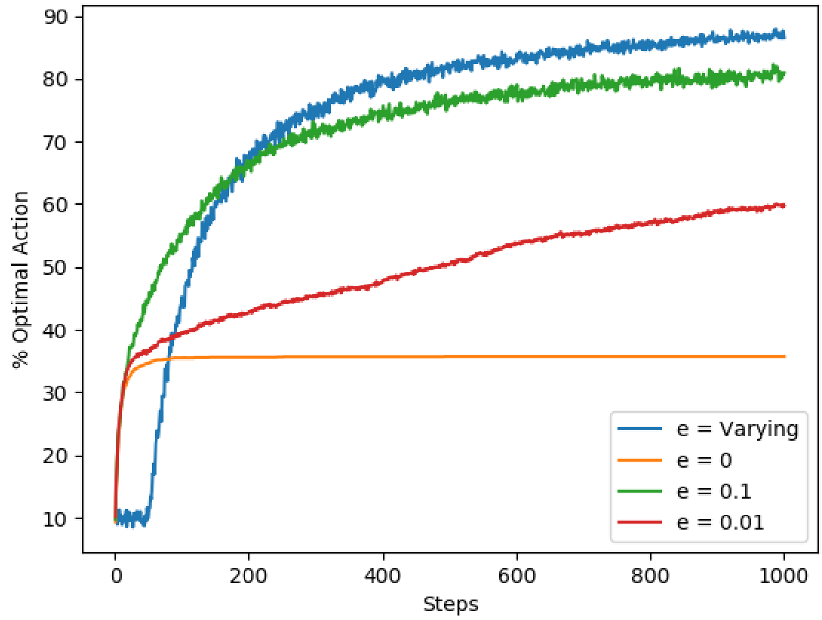

Fig 1.2: Optimal Action

The Varying Epsilon explores for the first 50-time steps and thus picks sub optimal actions in majority of the simulations, but in the long run it outperforms the other choices as it has sufficient knowledge about the estimates given it had exclusively explored during the initial stages. The Error in arms decreases in all cases, except when e = 0, since the first arm that it picks in any simulation, it picks that for the rest of the run, hence not exploring anything, for varying e, the arms are sufficiently and equally explored, hence error reduce is uniform across all arms

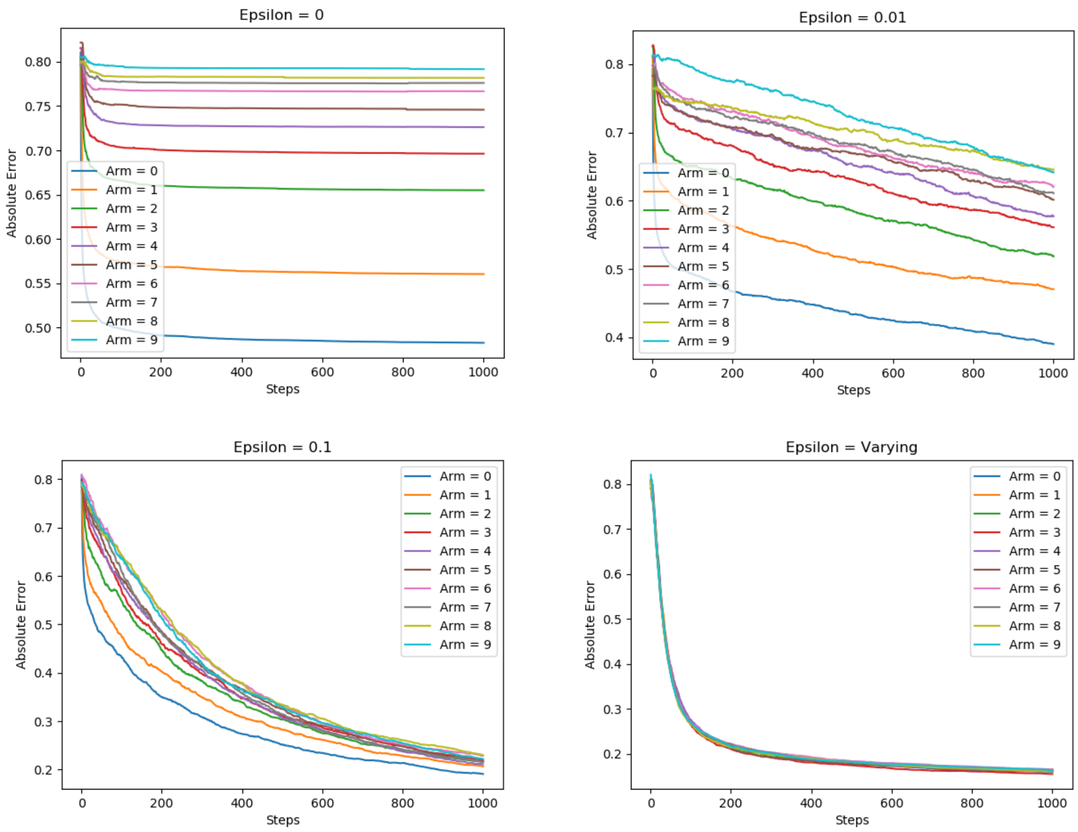

Fig 1.3: Error in Reward Estimate for different epsilon

Q2. As the variance across expected rewards increases, it takes slightly longer for the policies to converge to the Actual expected values, but still the earlier trend follows

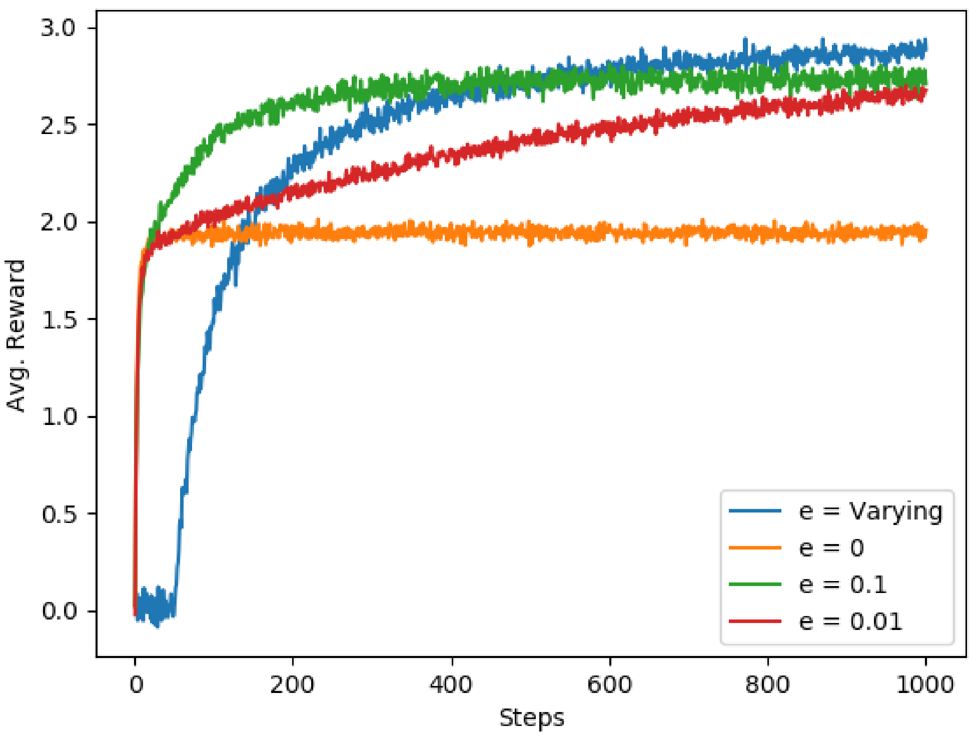

Fig 2.1: Average Reward

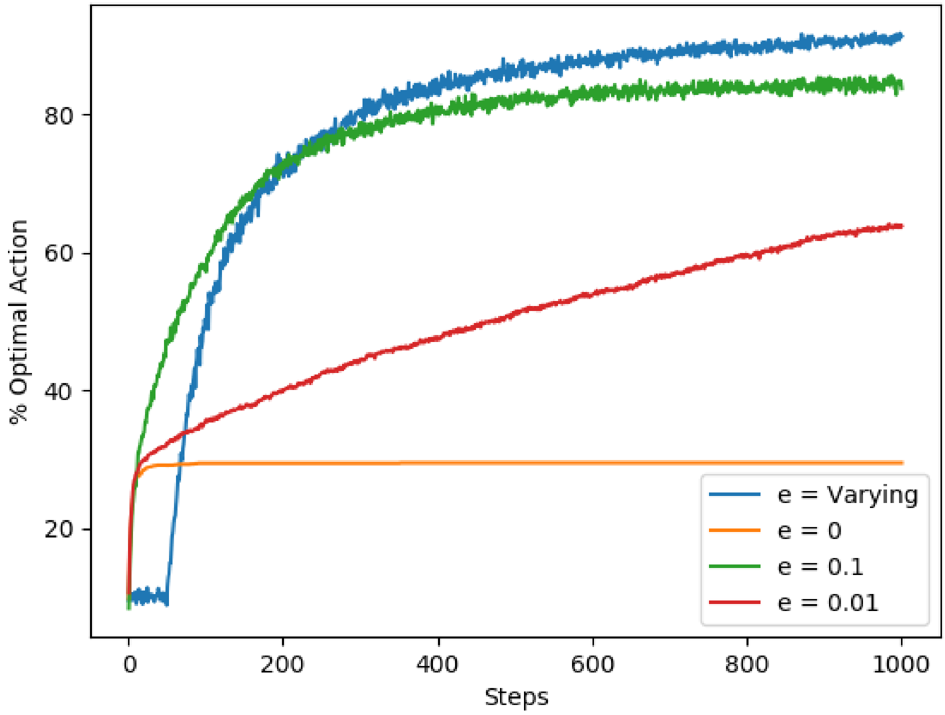

Fig 2.2: Optimal Action

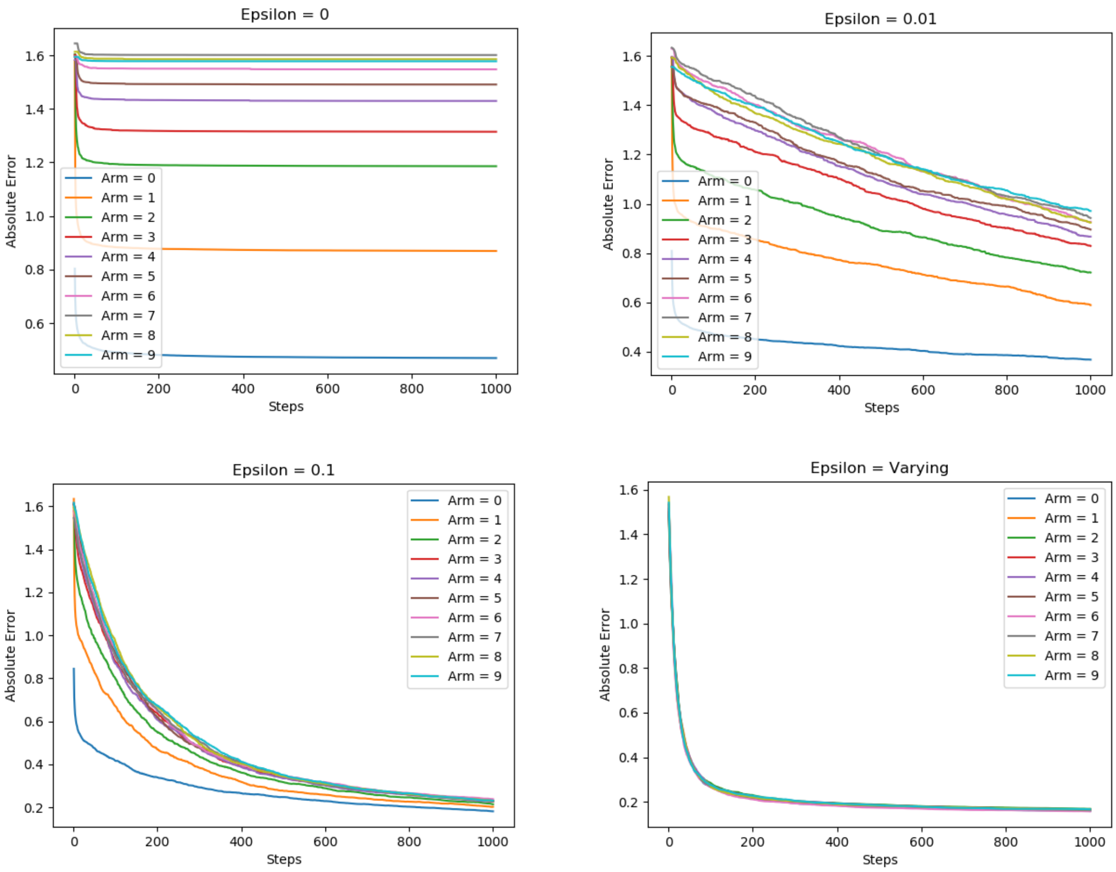

Fig 2.3: Error in Reward Estimate for different epsilon

Q3.

Q3. The estimates after a long time would converge to the actual Expected values.
at $t = \infty$, let estimates (= actual values) be

$$q_{opt}, q_1, q_2, \dots q_9$$

where q_{opt} is reward for optimal arm
$q_i \ \forall i \neq opt$ is reward for sub optimal arms

$$E(R) = E(R \mid A = \text{greedy}) \cdot P(\text{greedy})$$

$$+ E(R \mid A = \text{explore}) P(\text{explore})$$

$$= (1-\varepsilon) \cdot q_{opt} + \sum_{i \in A}^{9} \varepsilon \cdot \frac{1}{10} \cdot q_i$$

$$= (1-\varepsilon) q_{opt} + \varepsilon \cdot \frac{(q_1 + q_2 + \dots + q_9 + q_{opt})}{10}$$

let q_{avg} be the avg reward for all arms

$$E(R) = (1-\varepsilon) q_{opt} + \varepsilon \cdot q_{avg}$$

$$= q_{opt} + \varepsilon(q_{avg} - q_{opt})$$

clearly $q_{opt} \geq q_{avg}$

as $q_{opt} \geq q_i \quad \forall i \neq opt$

$$\frac{\sum_{i=1}^{10} q_{opt}}{10} \geq \frac{\sum_{i \in A} q_i}{10}$$

$$q_{opt} \geq q_{avg}$$

classmate

Date ______
Page ______

thus $E(R) = q_{opt} + \varepsilon(q_{avg} - q_{opt})$

as $q_{avg} - q_{opt} \leq 0$

the higher the ε, the lower the expected value.

Case ①: $\varepsilon = 0.1$

$$E_{0.1}(R) = q_{opt} + 0.1(q_{avg} - q_{opt})$$

Case ② $\varepsilon = 0.01$

$$E_{0.01}(R) = q_{opt} + 0.01(q_{avg} - q_{opt})$$

Case ③ $\varepsilon = \frac{50}{t} = 0$ as $t \to \infty$

$$E_{Conv}(R) = q_{opt}$$

Case ④ $\varepsilon = 0$

we cannot assume that estimates converge to actual values as we always pick ~~randomly~~ the only arm that had a positive reward

so the Expected value is q_{avg}.

Clearly

$$E_{Conv} > E_{0.01} > E_{0.1}$$

Thus the best choice is E_{Conv} with diff:

$0.01(q_{opt} - q_{avg})$ & $0.1(q_{opt} - q_{avg})$ for $\varepsilon = 0.01$ & 0.1

Next we have $\epsilon = 0.01$

with diff $0.09\ (q_{opt} - q_{avg})$

Q4.

Q4. Sample mean is not influenced by initial choice of $Q_1(a)$

Incremental update eq^n :-

(1) — $$Q_{n+1} = \left(\frac{n-1}{n}\right) Q_n + \frac{1}{n} R_n$$

(2) — $$Q_n = \left(\frac{n-2}{n-1}\right) Q_{n-1} + \frac{1}{n-1} R_{n-1}$$

putting (2) in (1)

(3) — $$Q_{n+1} = \left(\frac{n-2}{n}\right) Q_{n-1} + \frac{1}{n}\left(R_n + R_{n-1}\right)$$

(4) — $$Q_{n-1} = \left(\frac{n-3}{n-2}\right) Q_{n-2} + \frac{1}{n-2} R_{n-2}$$

putting (4) in (3)

$$Q_{n+1} = \left(\frac{n-3}{n}\right) Q_{n-2} + \frac{1}{n}\left(R_n + R_{n-1} + R_{n-2}\right)$$

in General

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

put $n-k=1$

$$Q_{n+1} = \left(\frac{0}{n}\right) Q_1 + \frac{1}{n} \sum_{i=1}^{n} R_i$$

$$= \sum_{i=1}^{n} \frac{R_i}{n}$$

which is independent of choice of Q_1

when using constant step parameter α

(1) — $Q_{n+1} = Q_n + \alpha(R_n - Q_n)$

(2) — $Q_n = Q_{n-1} + \alpha(R_{n-1} - Q_{n-1})$

(3) $Q_{n-1} = Q_{n-2} + \alpha(R_{n-2} - Q_{n-2})$

put (2) in (1)

(1.1) $Q_{n+1} = Q_{n-1} + \alpha(R_{n-1} - Q_{n-1})$
$+ \alpha(R_n - Q_{n-1} + \alpha(R_{n-1} - Q_{n-1}))$

$\vdots$

~~put (3) in (1.1)~~

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

$\vdots$

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

clearly Q_{n+1} is a function of Q_1

Also, the contribution of Q_1 to Q_{n+1} is given by

$$(1-\alpha)^n Q_1$$

Now consider two α : α_1, α_2 $0 \leq \alpha_i \leq 1$

where $\alpha_1 < \alpha_2$

$\Rightarrow 1-\alpha_1 > 1-\alpha_2$

$\Rightarrow (1-\alpha_1)^n > (1-\alpha_2)^n$

$(1-\alpha_1)^n Q_1 > (1-\alpha_2)^n Q_1$

$\Rightarrow$ For a smaller α, the influence of initial values is larger.

larger by :- $\left((1-\alpha_1)^n - (1-\alpha_2)^n\right) Q_1$

(ii) For complete eliminating influence of Q_1 on Q_n we can set $\alpha = 1$, thus

eqn becomes : $Q_{n+1} = R_n$

This although eliminates Q_1, but is a poor estimate for rewards.

Q5.

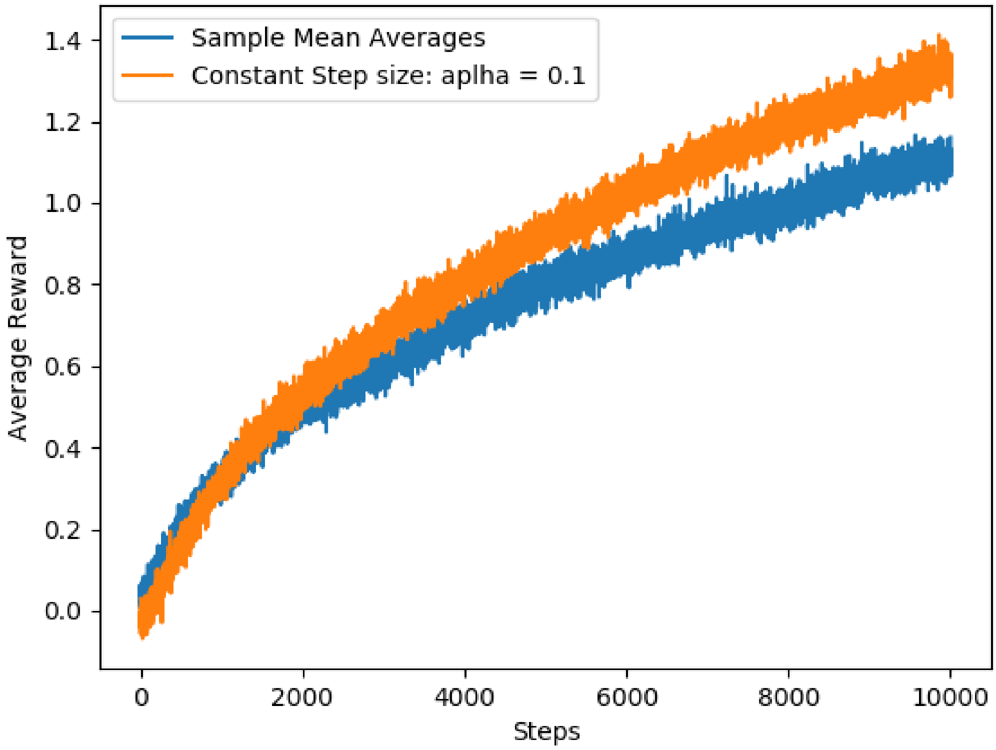

Fig 5.1: Average Reward

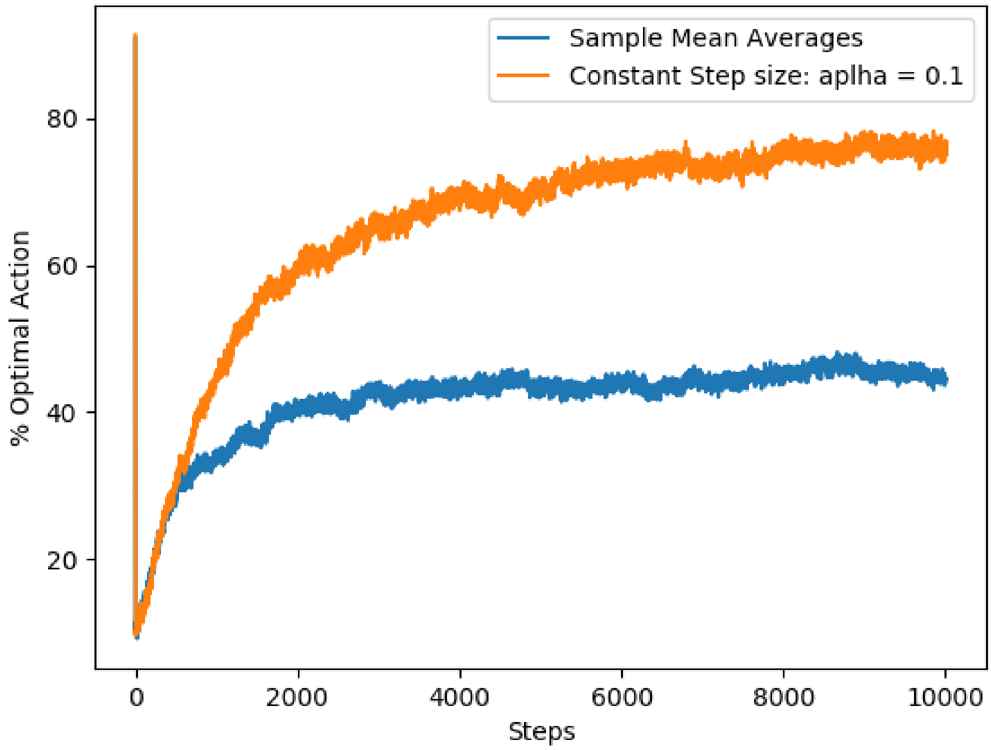

Fig 5.2: Optimal Action

Q6.

The spikes at the 11th time step are a result of the following: Since $N_t(a)$ for all a is zero, thus all arms are maximising, and whenever an arm is chosen, $N_t(a)$ for that arm a, becomes 1, i.e. that arm no longer remains maximising whilst there is some arm that yet hasn't been picked even once, so for the first ten time steps, all the ten arms are explored randomly once, at the 11th time step, the second term for each arm is same, so the agent picks the arm with the maximum estimate value Q_t, this happens across all 2000 runs, thus the average reward is higher, on the 12th time step, each of the 2000 runs take different independent arms, so the average reward drops, thus we obtain a distinct spike at the 11th time step

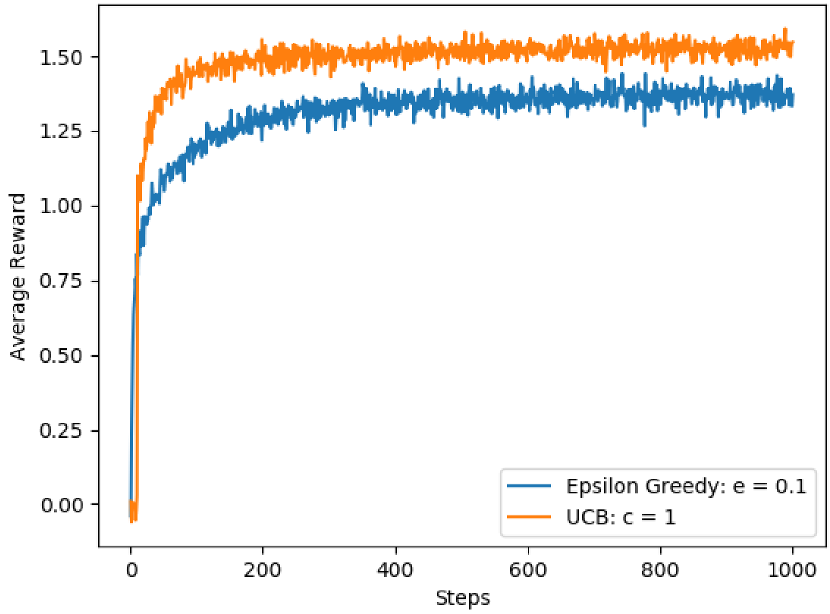

Fig 6.1: c = 1

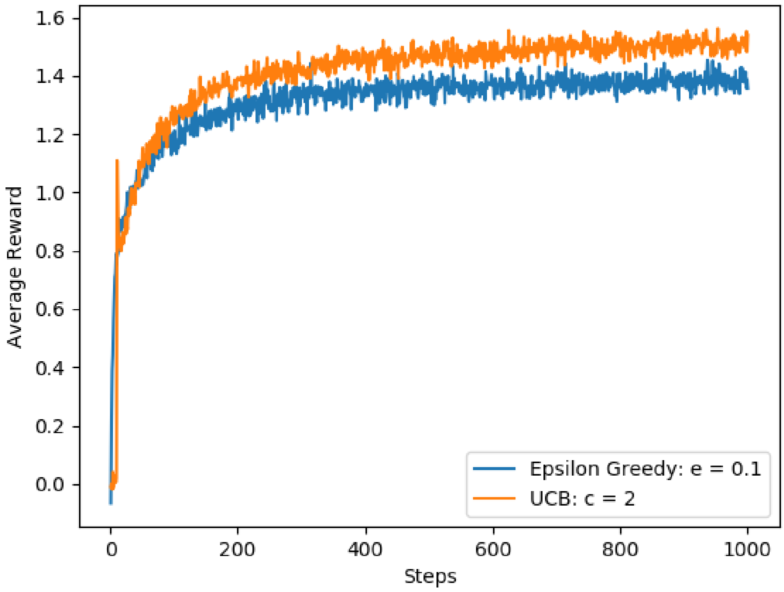

Fig 6.2: c = 2

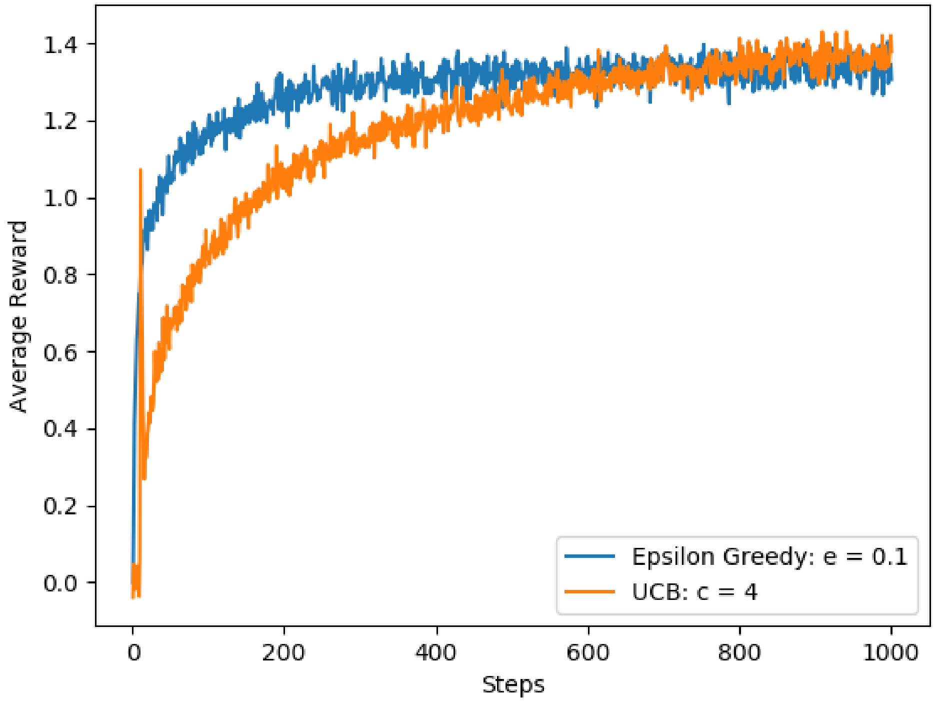

Fig 6.3: c = 4

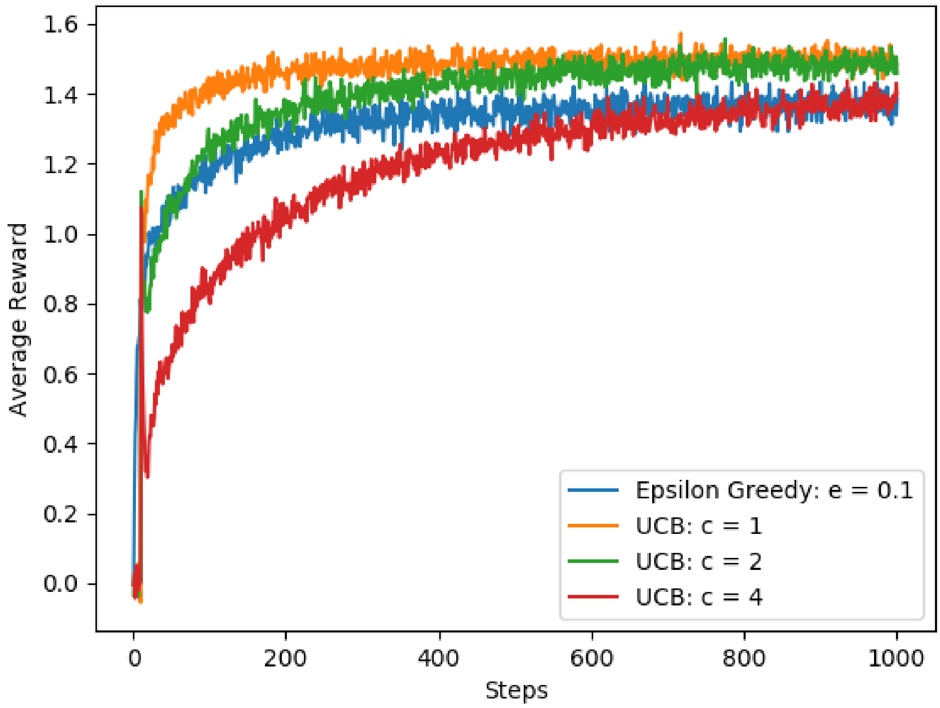

Fig 6.4: Combined

Q7.

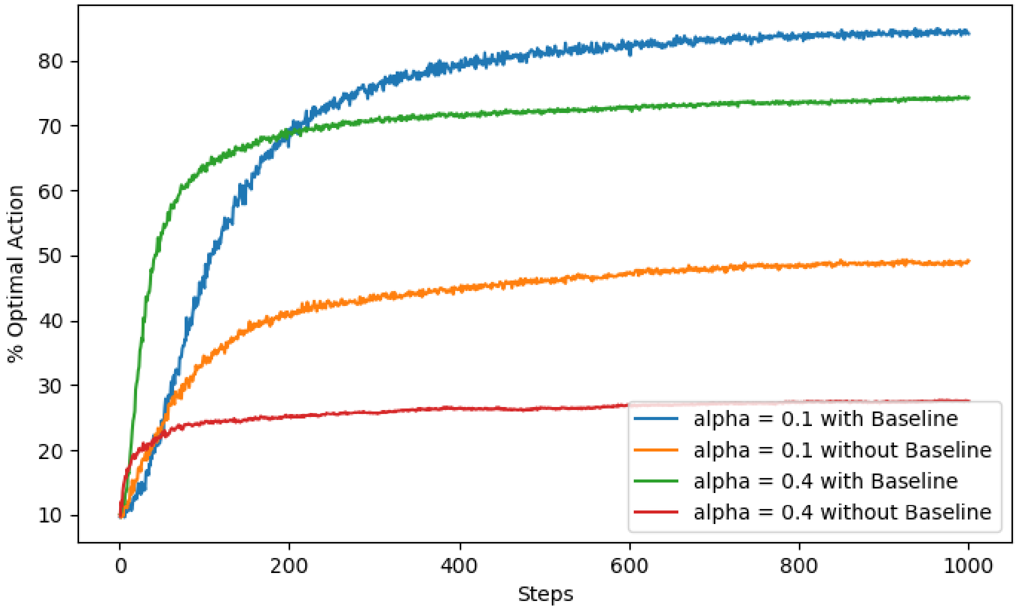

Fig 7.1: Gradient Bandit with alpha = 0.1, 0.4 with/without baseline