K - Armed Bandit

Goal

Goal of this assignment to study non associative, evaluative feedback problem K-Armed Bandit and compare several evaluative feedback techniques like greedy, ϵ-greedy methods, Optimistic Initial Value method on stationary and non-stationary environment.

Introduction

As per [1] K-Armed Bandit problem is defined as follows: Agent is faced repeatedly with a choice among k different options/actions. After each choice it receives a numerical reward chosen from a stationary probability distribution that depends on the action you selected. We compare standard action value and optimistic initial value methods on this problem. These techniques differ in the way they balance exploration and exploitation and hence provide valuable insights about stationary and non stationary environment.

Methodology

Approach

Epsilon Greedy Method:

In this method agent updates its initial estimates of actions on the basis of received rewards and balances exploration and exploitation by choosing exploratory action with ϵ-probability and optimal action rest of the time. Fig 1 shows the pseudo-code

```
Initialize, for $a = 1$ to $k$:
    $Q(a) \leftarrow 0$
    $N(a) \leftarrow 0$

Repeat forever:
    $A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$
    $R \leftarrow bandit(A)$
    $N(A) \leftarrow N(A) + 1$
    $Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\big[R - Q(A)\big]$
```

Fig 1 Epsilon Greedy Agent with sample weighting

Greedy Method:

In greedy method agent never explores the environment and always selects optimal estimate action.

Optimistic Initial Value Method:

This is an ϵ-greedy method in which action estimates are set optimistically high.

For all these methods we have 2 variants: sample weighting and exponential recency weighting. Sample weighting is more suitable for stationary environment whereas exponential recency weighting is more suitable for non stationary environments since it weights recent actions higher than older actions.

Implementation

- **Framework:** We have used Open AI Gym framework to implement the above approaches and K-Armed Bandit environment.
- **Testbed:** Created a testbed of 2000 instances of K-Armed Bandit as described in [1] and show below

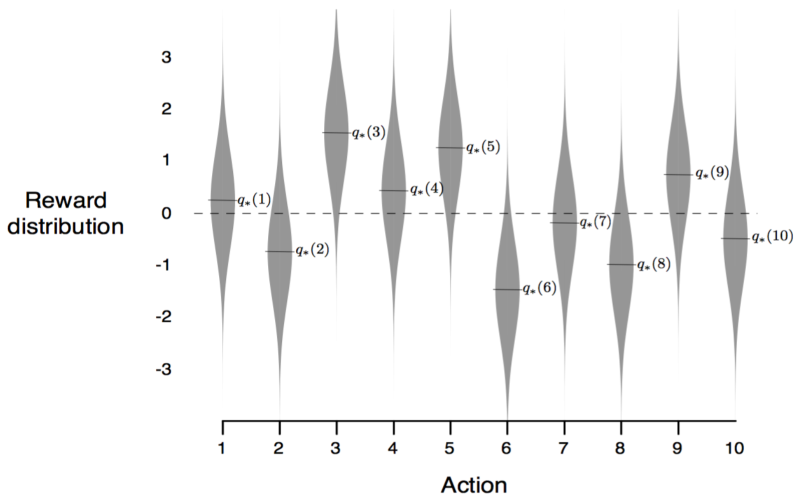


Fig 2 10-armed testbed with actions drawn from Gaussian with mean 0 and variance 1 & rewards drawn from Gaussian with variance 1 around actions

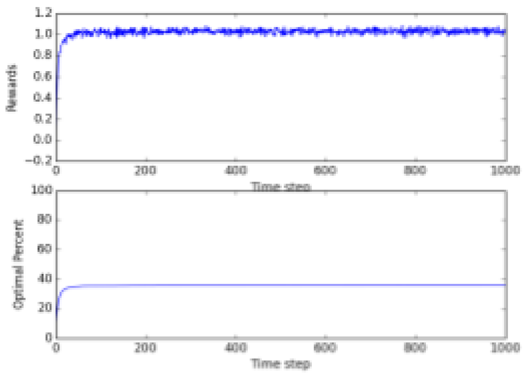
Experiments & Results

We have used the following values of parameters in all the experiments unless mentioned otherwise:

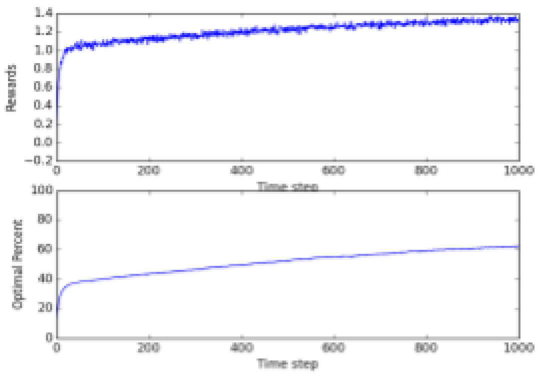
- ϵ = 0.1
- weighting scheme = sample weighting
- α = 0.1 for recency weighting
- Action estimates drawn from Gaussian with mean 0 and variance 1
- Reward variances = 1
- Optimistic Initial Value = 5

E1: Comparison of all techniques for Stationary Environment

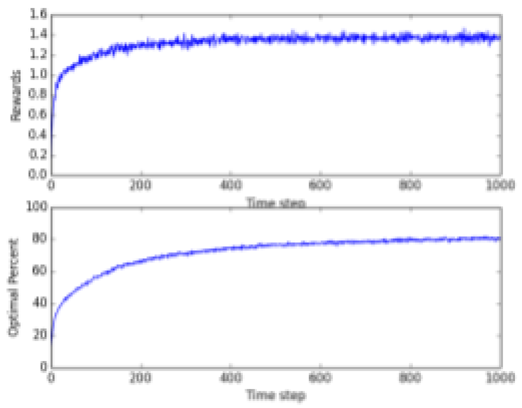
In this experiment we compare all the approaches on the testbed with stationary environment and reproduce the plots in Ch 2 of textbook [1] We observe that as expected for greedy agent (ϵ = 0) percent optimal action is lower as compared to other methods. We also observe that agent with higher ϵ (0.1) starts taking optimal actions earlier than agent with lower epsilon (0.01) since it explores more and finds optimal action earlier. Similarly agent with recency weighting (α = 0.1) takes more time as can be seen in beginning part of optimal percent plot (right bottom) but catches up later. Important thing to note in this experiment that all agents except greedy agent explore and hence were able to identify optimal action and had rewards hovering around 1.4



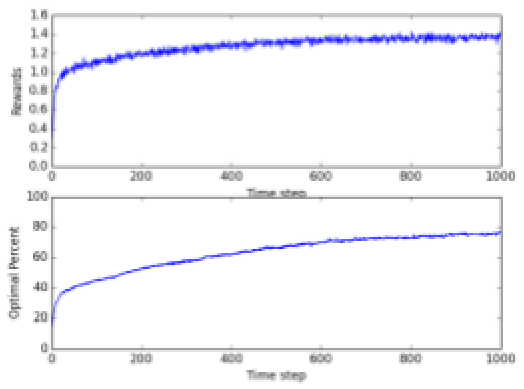
$\epsilon = 0$



$\epsilon = 0.01$



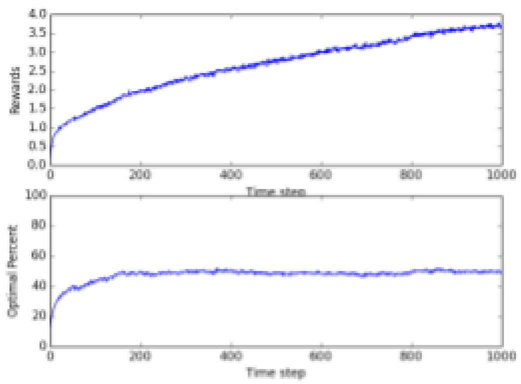
$\epsilon = 0.1$



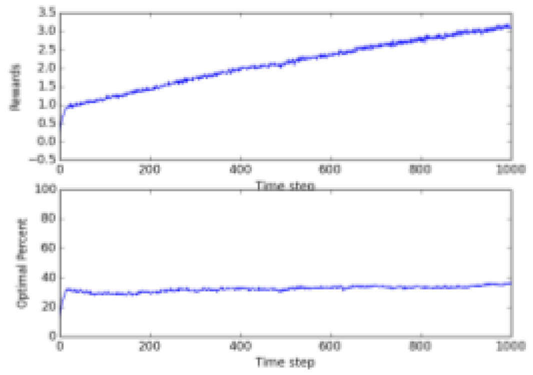
$\epsilon = 0.1$, RecencyWeighting $\alpha = 0.1$

E2: Comparison of all techniques for Non Stationary Environment

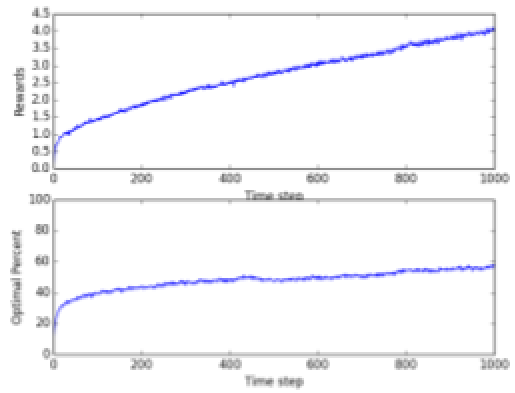
Through this experiment we demonstrate the difficulties of sample average method for non stationary environments. In this experiment the $q_*(a)$ start equal and then take random walk (with random walk var 0.1) as described in exercise 2.3 of Ch 2 in [1]. If we look at optimal action percent plots for 4 methods we observe that sample average and recency weighting with $\alpha = 0.01$ struggle for non stationary environment. Recency weighting with $\alpha = 0.1$ performs better with optimal percent reaching around 55% and reward around 4. Recency weighting with $\alpha = 0.9$ performs best with optimal percent reaching around 70% and reward around 4.5. This shows that in non stationary environments recency weighting performs much better than sample average as it adapts better to dynamic environment.



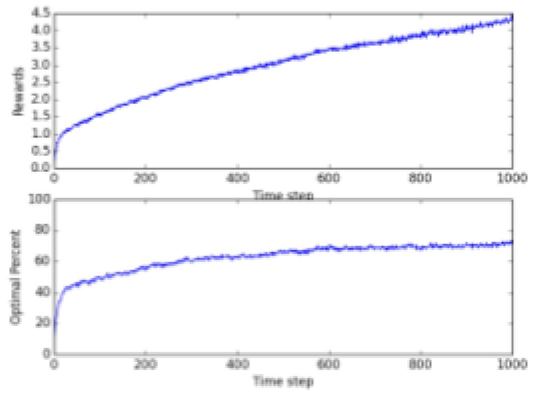
Sample Average



Recency Weighting, α = 0.01



Recency Weighting, α = 0.1

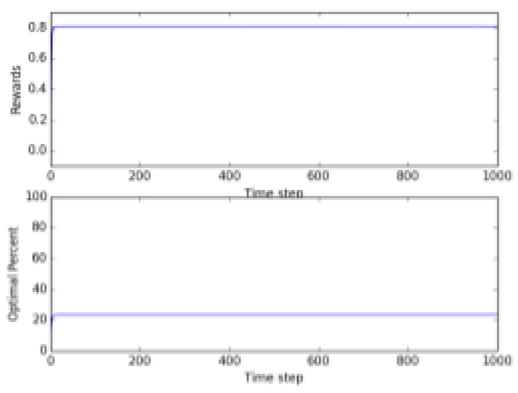


Recency Weighting, α = 0.9

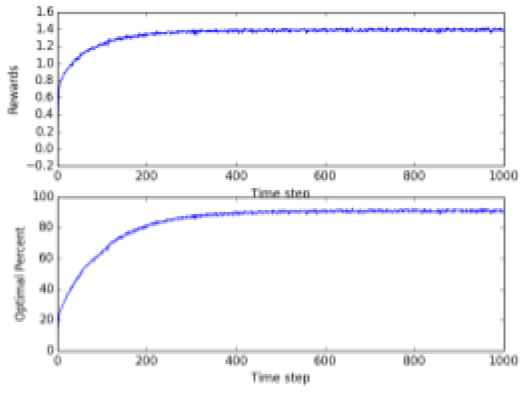
E3: Low vs High Reward Variance

In this experiment we compare performance of greedy and ϵ-greedy methods for environments with low and high reward variance. We observe that if reward variance is very low then greedy agent can perform well if we use optimistic initial value to force it to explore all actions at least once. Once it has done this it will dominate epsilon greedy which will keep exploring even though reward variance is very low.

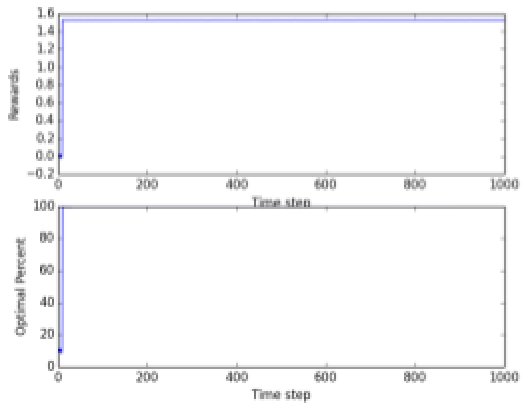
In high reward variance environment greedy agent performs bad even with optimistic initial value. As expected ϵ-greedy agent performs better (mean reward 1.4 v/s 1 for greedy agent and optimal action percent of 50 v/s 30 for greedy agent)



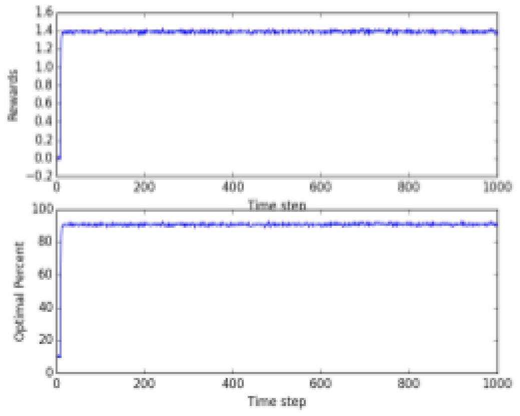
ϵ = 0, Reward Var = 0.1



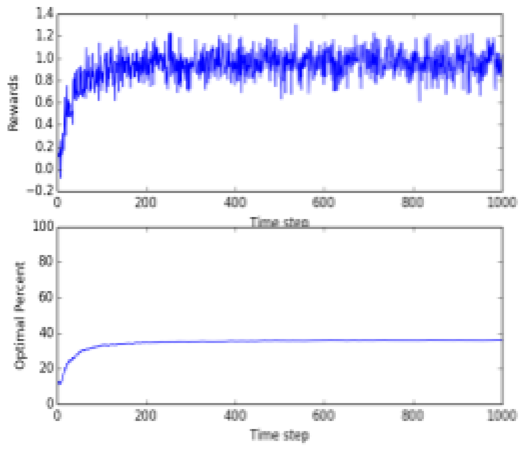
ϵ = 0.1, Reward Var = 0.1



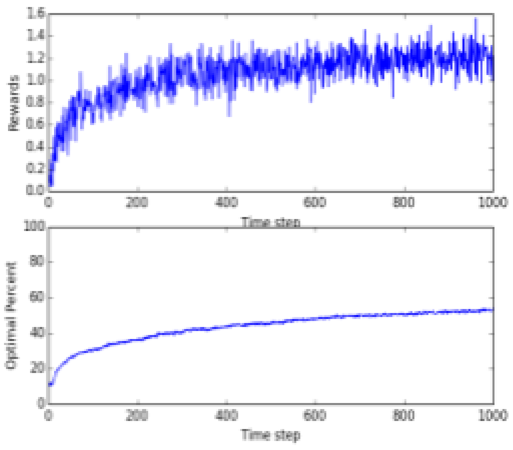
ϵ = 0, Initial Value 5, Reward Var = 0.1



ϵ = 0.1, Initial Value 5, Reward Var = 0.1



ϵ = 0, Initial Value = 5, Reward Var = 5



ϵ = 0.1, Initial Value = 5, Reward Var = 5

References

[1] Reinforcement Learning: An Introduction