Assignment 1 - Foundations of Intelligent & Learning Agents

Kartik Gokhale

Sept 2022

Contents

1 Task 1

1.1 UCB

In the case of Upper Confidence Bounds, we see the regret is sub-linear and thus, it performs better than the epsilon-greedy algorithms in terms of optimising regret. Noticing the constant, the algorithm generates a regret of roughly 1400.

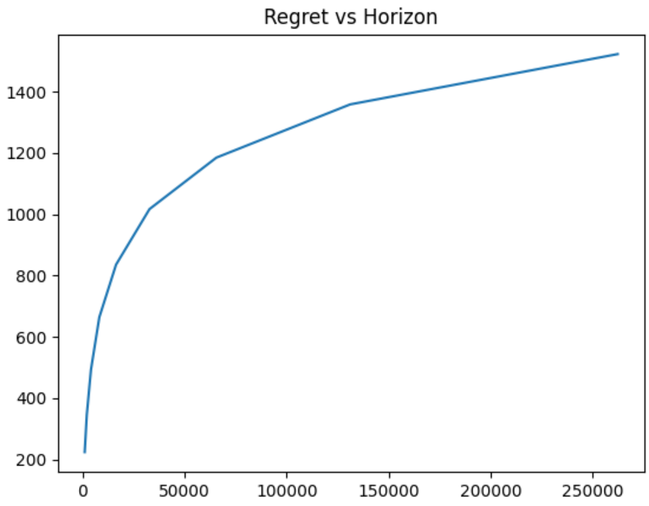

Figure 1: UCB

1.2 KL-UCB

In the case of KL - Upper Confidence Bounds, we see the regret is also sub-linear and thus, it also performs better than the epsilon-greedy algorithms in terms of optimising regret. Noticing the constant, the algorithm generates a regret of roughly 220, which makes it a tighter bound on regret than the UCB algorithm.

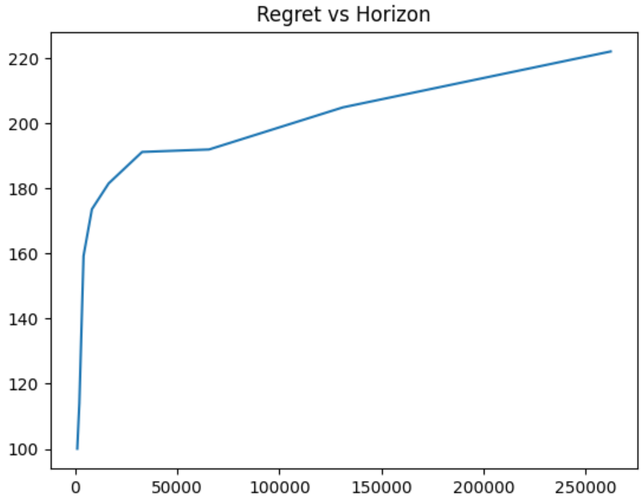

Figure 2: KL-UCB

1.3 Thompson Sampling

In the case of Thompson, we again see the regret is also sub-linear and thus, it also performs better than the epsilon-greedy algorithms in terms of optimising regret. Noticing the constant, the algorithm generates a regret of roughly 130, which makes it a tighter bound on regret than the UCB and KL-UCB algorithms.

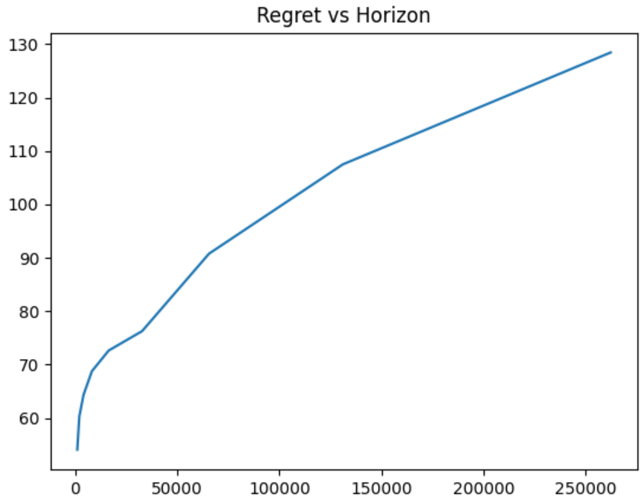

Figure 3: Thompson Sampling

2 Task 2

My strategy in this problem treats batched sampling as a limitation on policy change. This means that the arms to be sampled are pre-determined for an entire batch and any and all beliefs about the arms can be updated only at the end of a batch.

- Each pull from batch B_i is an independent pull from a thompson sampling with beliefs based on results from all batches before batch B_i
- The results(rewards) from all pulls are used to update the beliefs about the k-arms

The plot comes to be linearly increasing with batch size as the increase in batch size reduces the ability to update the beliefs about arms. In the extreme case, when batch size equals horizon, it will be a truly random choice of arms and thus, maximal possible regret.

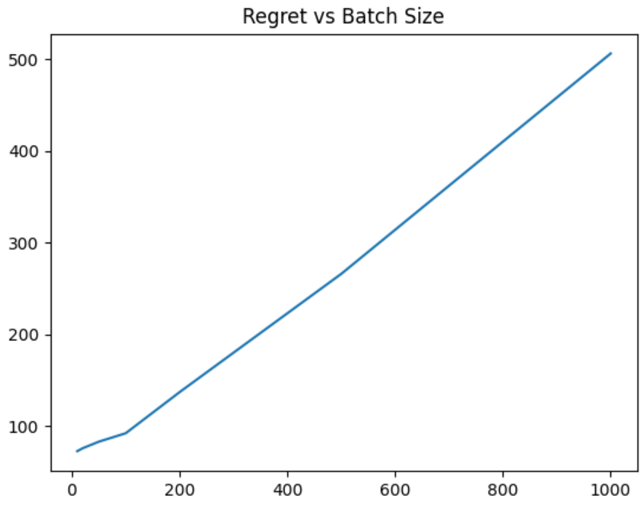

Figure 4: Batched Sampling

3 Task 3

For this task, I maintained my belief about each of the arms based on successes and failures. As long as my belief about an arm was in the top 5%, I kept sampling it. Else, I would sample another arm. If a sub-optimal arm had high reward, I would still be performing above sampling each arm once and the moment an arm would show it's sub-optimality, we can move to another arm.

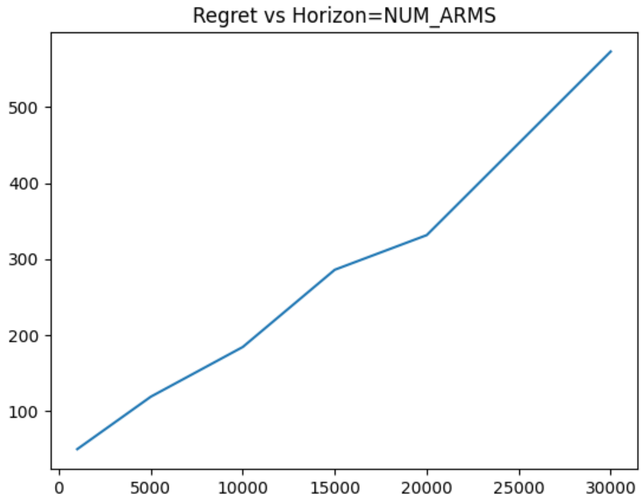

Figure 5: Finite Horizon