Reinforcement Learning and Dynamic Optimization

Report

Assignment 1 – Recommending News Articles to Unknown Users.

Student name: Asterinos Karalis

Student AM: 2020030107

Introduction

The goal of this assignment is to implement the UCB algorithm for the click-through rate problem in Python, prove an upper bound for the regret considering the problem's characteristics and plot it alongside the actual cumulative regret generated by the simulation for different horizon values.

Proving the upper bound for regret

For the current problem we are considering the Instance-Independent-Bound since each round is totally independent from one another. According to Slivkins's book:

Let's consider two sets of arms:

1. The set of arms i ∈ S, where 0< Δi ≤ ϵ
2. The set of arms i ∈ S′ (S's complement), where Δi> ϵ

Then we obtain our instance-independent bound assuming the clean event:

$$R(T) = \sum_{i=1}^{K} N(T)_i \, \Delta_i = \sum_{i \in S} N(T)_i \Delta_i \; + \sum_{i \in S'} N(T)_i \, \Delta_i \; \leq \; \epsilon \sum_{i \in S} N(T)_i + log(T) \sum_{i \in S'} \frac{1}{\Delta_i}$$

For the first part we know that $\sum_{i=1}^{K} N(T)_i \; = \; T$, therefore $\epsilon \sum_{i \in S} N(T)_i$ is bounded by ϵT.

Also, for the second part we know that since we are in S′ and the fact that there are at most K arms, we can bound the $log(T) \sum_{i \in S'} \frac{1}{\Delta_i}$ by log(T)K/ϵ.

To get the strictest upper bound, we use the Δ that minimizes second part of the R(T) bound.

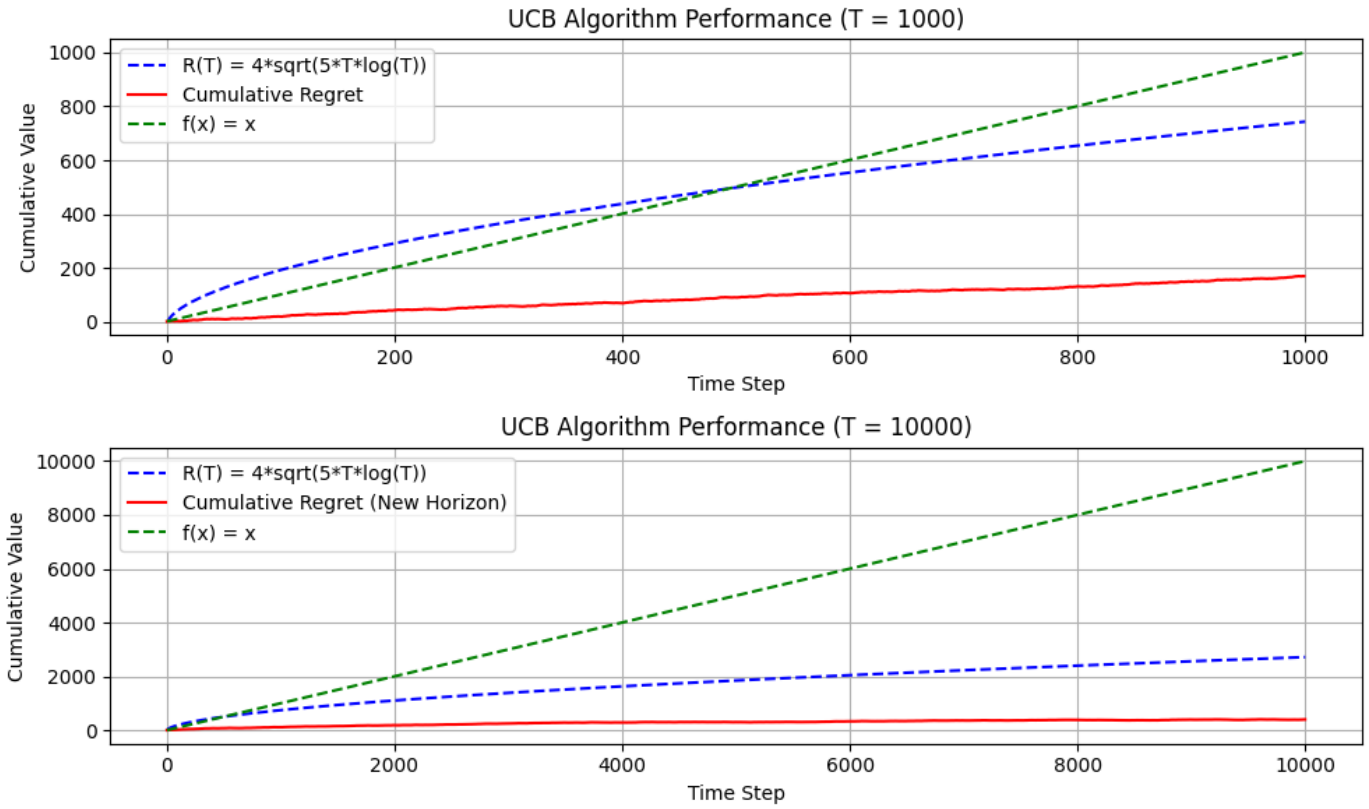
We do this by setting ϵT=log(T)K/ϵ which gives $\epsilon \; = \; \sqrt{\frac{Klog(T)}{T}}$. We now arrive at our final bound which is roughly $R(T) \leq \sqrt{KTlog(T)}$.

Considering that for our problem, our algorithm must learn about each user's taste in articles and therefore accumulate regret from each user until it can become efficient. So we end up gathering regret from each user and arrive at our new regret bound:

$R_{TOTAL} \leq U \; R(T) = \; U \sqrt{KTlog(T)}$ where U is the number of different users (in our case U=4)

Finaly we get the regret bound: $R(T) \leq 4\sqrt{5Tlog(T)}$

Simulation Results



As we can see from the results of the simulation, the cumulative regret is way below its theoretical limit and becomes a lot smoother as the horizon increases (seems to be completely flat).

For a shorter horizon we observe that in the worst cases we could get a regret that can be worse than linear for nearly the first half of the simulation. However, when we have a far larger horizon we are far below the linear line.

Overall, the cumulative regret of the simulation remains under the upper bound which is to be expected as the algorithm gathers information about each user type and makes mostly optimal choices.

Bibliography:

https://kfoofw.github.io/bandit-theory-ucb-analysis/

https://people.eecs.berkeley.edu/~jiantao/2902021spring/scribe/EE290_Lecture_04.pdf

https://arxiv.org/abs/1904.07272