**Technical University of Crete**

**School of Electrical and Computer Engineering**

Course: **Reinforcement Learning and Dynamic Optimization**

Assignment 1

Angelopoulos Dimitris - 2020030038

In this exercise our goal is to develop a modified version of the UCB algorithm to recommend one of $K = 5$ articles to a given class of users. Each user can be female or male and over or under 25 years old; that means that we have $|U| = 4$ types of users. Different types of users have different probabilities $p_i$ to click article $i$.

We will suppose that every user visits our website randomly in an IID manner. The modified UCB algorithm will be the same with the default one parameterized by the user type, i.e we will randomly generate a user type and use it to modify our environment.

First of all we will find an upper bound for the regret. Using the law of total expectations we get :

$$\mathbb{E}[R(T)] = \sum_{u \in \mathbb{U}} P(U = u)\mathbb{E}[R(T)|U = u]$$

Where $\mathbb{E}[R(T)|U = u]$ is the regret for a given user.

As we mentioned the modified UCB algorithm will be the default one for a given user meaning that the following holds.

$$\mathbb{E}[R(T)|U = u] = P(\text{Good})\sum_{i=1}^{K} N_{i,u}(t)\Delta_{i,u} + P(\text{Bad})\sum_{i=1}^{K} N_{i,u}(t)\Delta_{i,u}$$

Using Hoeffding's inequality we get that $P(\text{Bad}) \leq KT^{-3}$ and $P(\text{Good}) \to 1$. Also $N_{i,u}(t)\Delta_{i,u} \leq T$. Thus the Bad event term vanishes as $T$ approaches infinity.

$$\left.\begin{matrix} \mathbb{E}[R(T)|U = u] \leq \sum_{i=1}^{K} N_{i,u}(t)\Delta_{i,u} \\ \Delta_{i,u} \leq 2\sqrt{\frac{2\log T}{N_{i,u}}} \end{matrix}\right\} \Rightarrow \mathbb{E}[R(T)|U = u] \leq \sum_{i=1}^{K} \sqrt{N_{i,u}(t) 8 \log T}$$

The maximum amount of times an article is presented is $T$, thus $N_{i,u}(t) \leq T$. This yields:

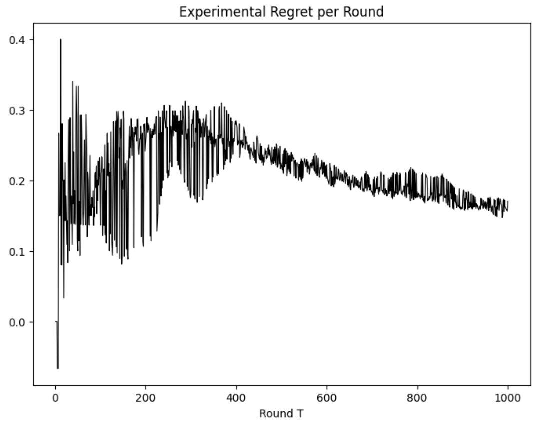
$$\mathbb{E}[R(T)|U = u] \leq K\sqrt{8T\log T}$$

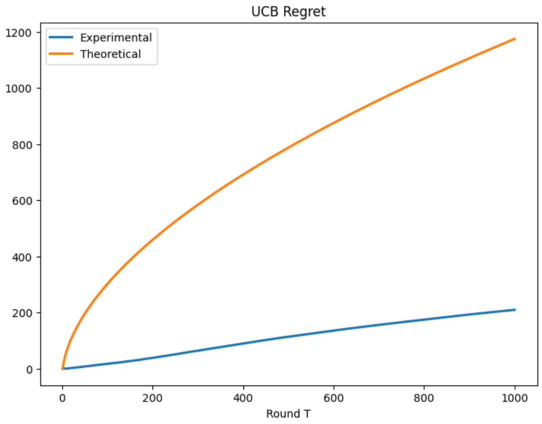
Substituting to the first equation we get the following upper bound of the regret.

$$\mathbb{E}[R(T)] \leq K \sum_{u \in \mathbb{U}} P(U = u)\sqrt{8T \log T} = K\sqrt{8T \log T}$$

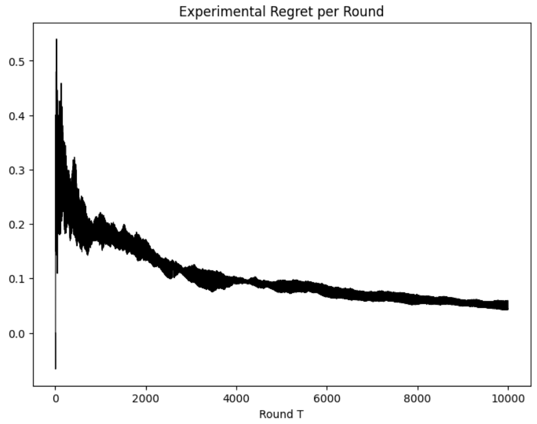
After running our program for time horizon $T$ we plot the experimental regret per round $t$ and $\mathbb{E}[R(T)]$.

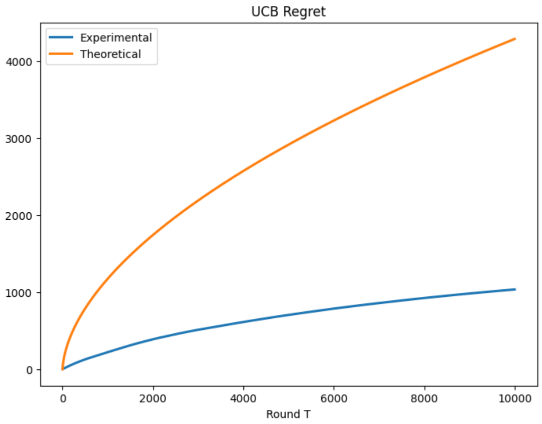
First of all we set $T = 1000$ and we get the following figures :





For $T = 10000$ we get :





In both cases we can verify that the regret per round $t$ approaches 0 as the time horizon gets larger. We also observe that the regret values are noisy; that is because the UCB algorithm will always choose to present articles other than the one with the larger mean.

In the expected regret case we can verify that indeed our theoretical regret upper bounds the experimental one. For $T = 1000$ the experimental regret seems linear; that is because the time horizon is not large enough and our algorithm tends to present different articles more frequently.

The Colab code for the assignment can be found in the following link.