# Reinforcement Learning and Dynamic Optimization (ΠΛΗ423/ΠΛΗ723)
# Assignment 1 : Recommending News Articles to Unknown Users

Nikolaos Angelidis

April 16, 2024

## 1 Introduction

The objective we were asked to accomplish for our first assignment was to make a model that selects an article for visiting user to our website, that is most likely to be clicked on and read more thoroughly (click-through rate). More specifically we have 5 news articles to choose from and different classes of users are female over 25, male over 25, male under 25 and female under 25.

## 2 Measurement plots

### 2.1 Plots for T=1000

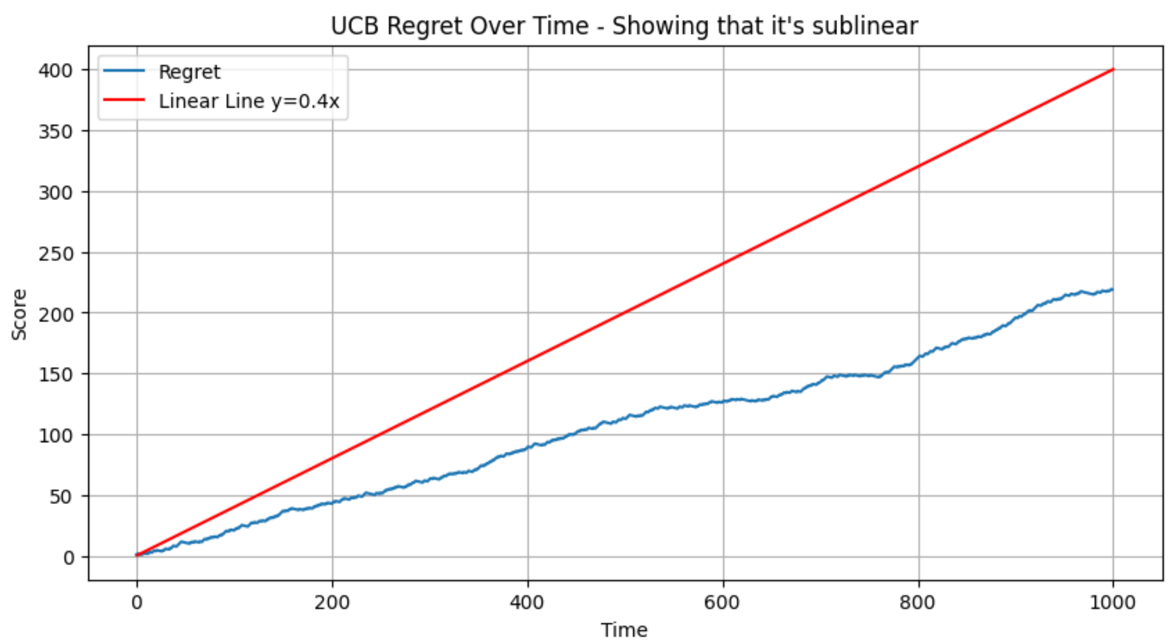


Figure 1: Sub-linear regret of the UCB Algorithm for T = 1000

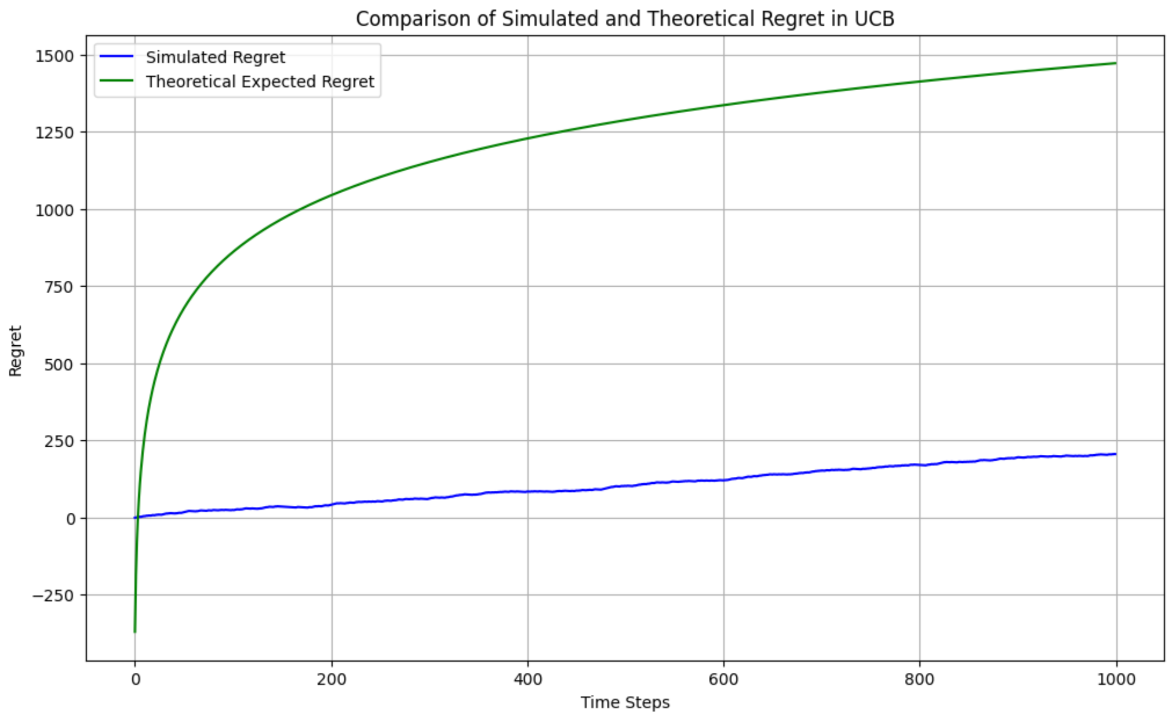


Figure 2: Comparison of theoretical and simulated regret for T = 1000

## 2.2 Plots for T=10000

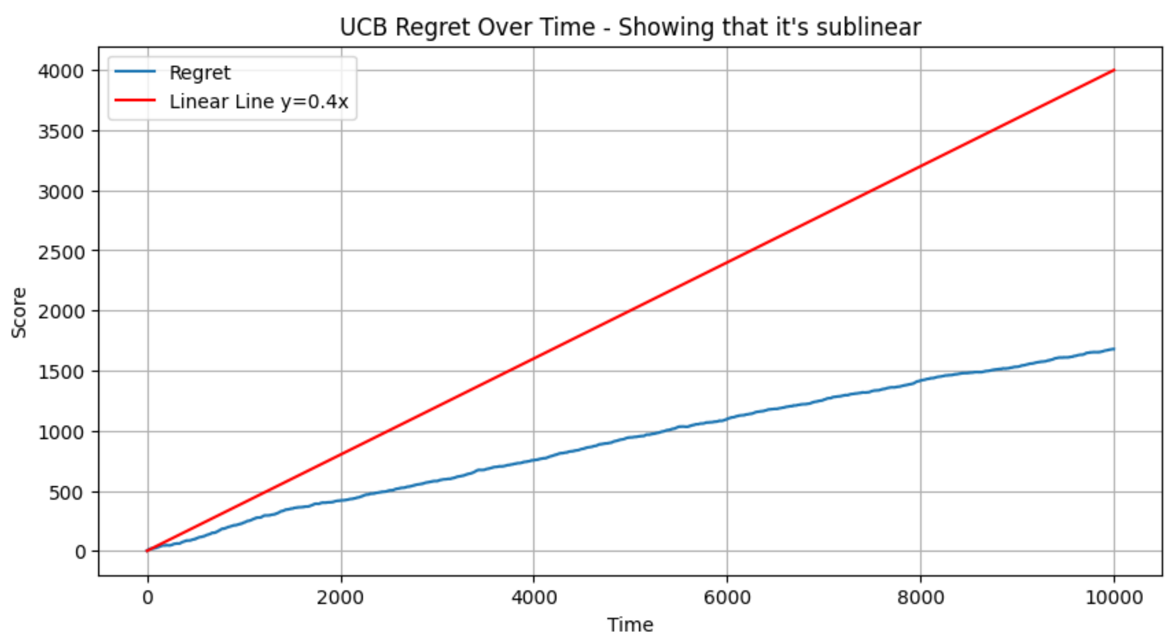


Figure 3: Sub-linear regret of the UCB Algorithm for T = 10000


Figure 4: Comparison of theoretical and simulated regret for T = 10000

### 2.3 Comparison of different horizon sizes

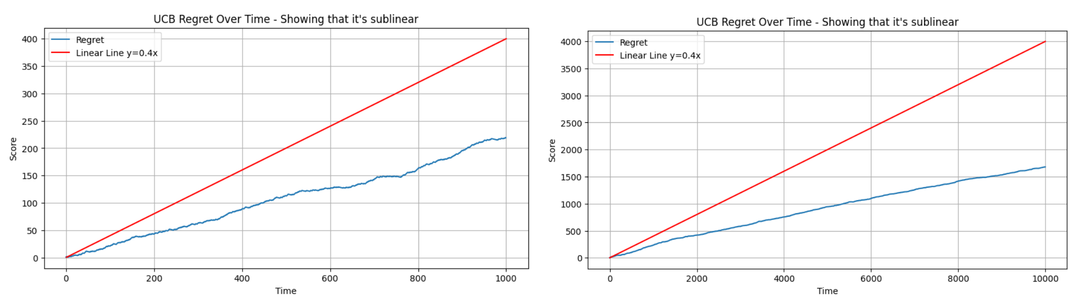


Figure 5: Comparison of simulated regret for T=1000 and T=10000

The two plots provide insightful views into the performance of our UCB algorithm over different lengths of operation (T=1000, T=10000). Both demonstrate a sub-linear growth pattern, which is indicative of the algorithm's effectiveness in improving over time. As the algorithm continues to run for a longer duration (larger horizon T), it appears to settle into a more consistent improvement pattern, with reduced fluctuations and steadier progress.

## 3 Theoretical Expected Regret Derivation

$$E[R(T)] = \sum_{u=1}^{U}\sum_{t=1}^{T}\Delta_{u,i}X_{u,i,t} = \sum_{u=1}^{U}\sum_{i=1}^{K}N_{u,i}(t)\Delta_{u,i} \tag{1}$$

$$|\hat{\mu}_{u,i}(t) - \mu_{u,i}| \leq \sqrt{\frac{2log(T/U)}{N_{u,i}(t)}}$$

,because each user appears T/U over horizon T on uniformly distributed.

$$P(Bad) = P(\exists u, i, t : |\hat{\mu}_{u,i}(t) - \mu_{u,i}| > \sqrt{\frac{2log(T/U)}{N_{u,i}(t)}} \leq K \cdot U \cdot T \cdot T^{-4}$$

For user u, at round t arm i was played:

$$\mu_{u,i} + 2\sqrt{\frac{2log(T/U)}{N_{u,i}(t)}} \geq \hat{\mu}_{u,i} + \sqrt{\frac{2log(T/U)}{N_{u,i}(t)}} (since \mu_{u,i} + \sqrt{\frac{2log(T/U)}{N_{u,i}(t)}} \geq \hat{\mu}_{u,i})$$

$$\geq \hat{\mu}^*_{u,i} + \sqrt{\frac{2log(T/U)}{N^*_{u,i}(t)}} \geq \hat{\mu}^*_u$$

So we have:

$$\Delta_{u,i} \leq 2\sqrt{\frac{2log(T/U)}{N_{u,i}(t)}}]$$

We use the formula from the presentation to bound $N_{u,i}(t)$ , for every arm and user:

$$N_{u,i}(t) \leq \frac{8log(T/U)}{\Delta^2_{u,i}}$$

$$N_{u,i}(t) \cdot \Delta_{u,i} \leq \frac{8log(T/U)}{\Delta_{u,i}} \tag{2}$$

Consequently, from (1) we get:

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

Term 1:

$$P(Good) \cdot \sum_{u=1}^{U} \sum_{i=1}^{K} N_{u,i}(t)\Delta_{u,i}$$

Term 2:

$$P(Bad) \cdot \sum_{u=1}^{U} \sum_{i=1}^{K} N_{u,i}(t)\Delta_{u,i}$$

For Term 2, we know:

$$P(Bad) \leq K \cdot T^{-3} (N_{u,i}(t) \cdot \Delta_{u,i} \leq T)$$

$$P(Bad) \cdot \sum_{u=1}^{U} \sum_{i=1}^{K} N_{u,i}(t)\Delta_{u,i} \leq U \cdot K \cdot T^{-3} \cdot T = U \cdot K \cdot T^{-2} \text{ (As T grows, we can ignore)}$$

So (*) becomes:

$$E[R(T)] = \sum_{u=1}^{U} \sum_{i=1}^{K} N_{u,i}(t)\Delta_{u,i} \overset{(2)}{\leq} \sum_{u=1}^{U} \sum_{i=1}^{K} \frac{8log(T/U)}{\Delta_{u,i}}$$

$$E[R(T)] \leq \sum_{i=1}^{K} U \cdot \frac{8log(T/U)}{\Delta_i}$$

$$E[R(T)] \leq K \cdot U \cdot \frac{8log(T/U)}{\Delta_i} = 8 \cdot K \cdot U \cdot \frac{log(T/U)}{\Delta_i}$$