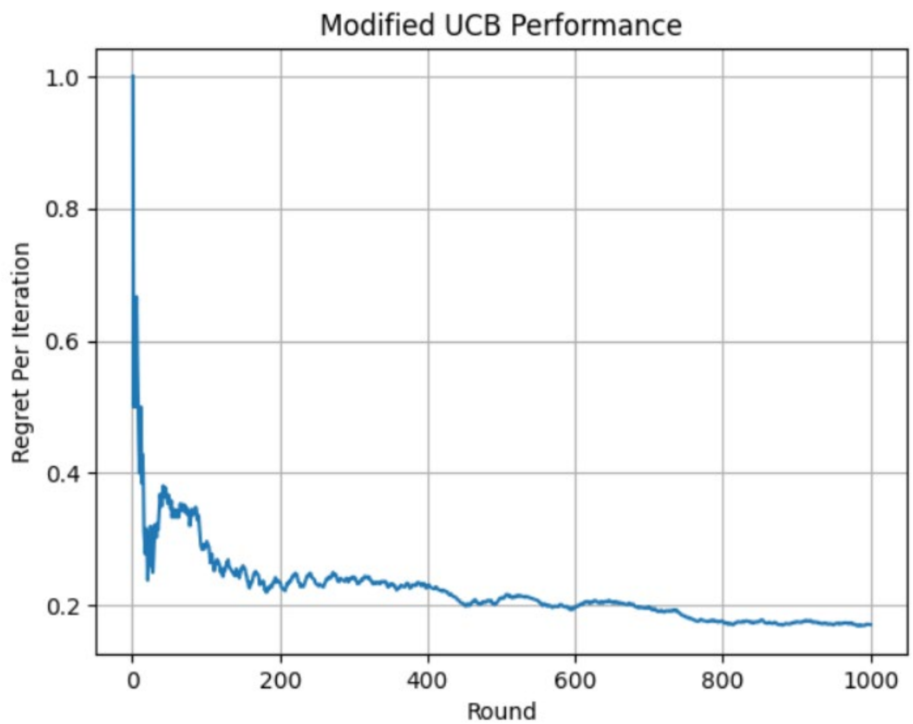
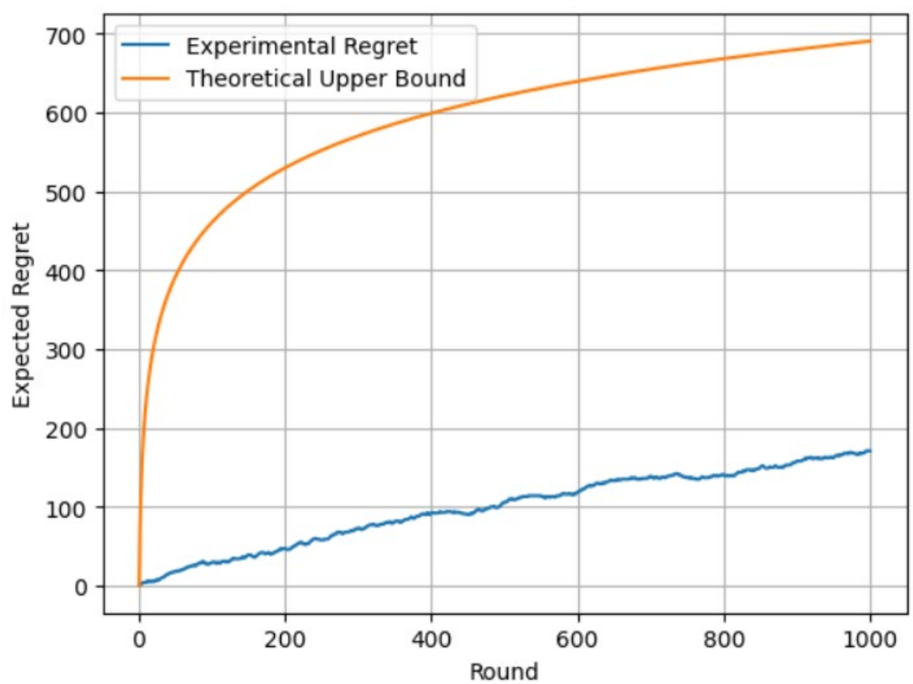
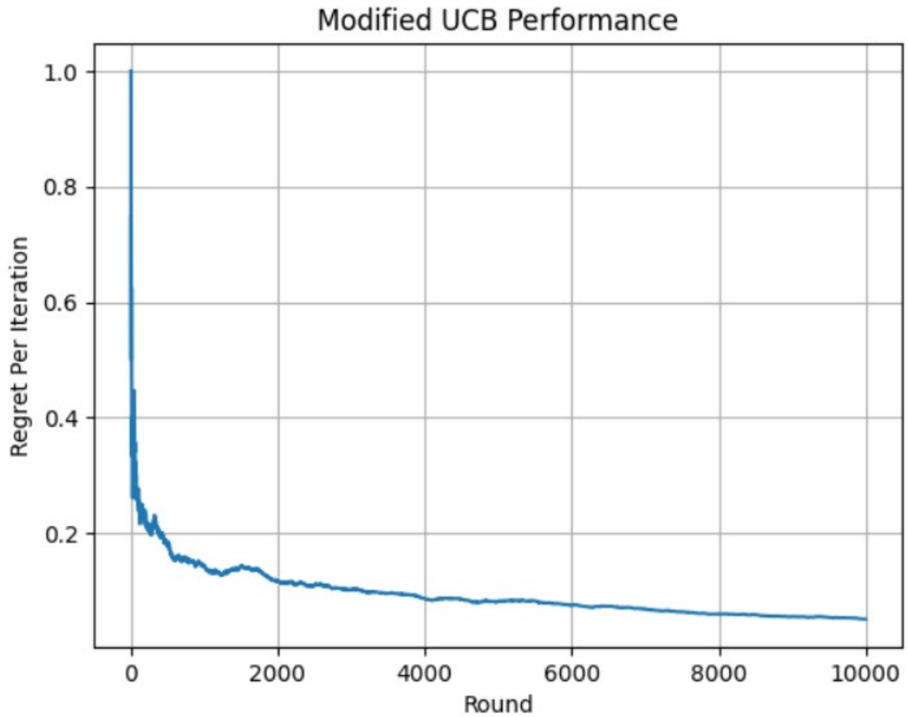
Technical University of Crete
School of Electrical and Computer Engineering
Reinforcement Learning and Dynamic Optimization

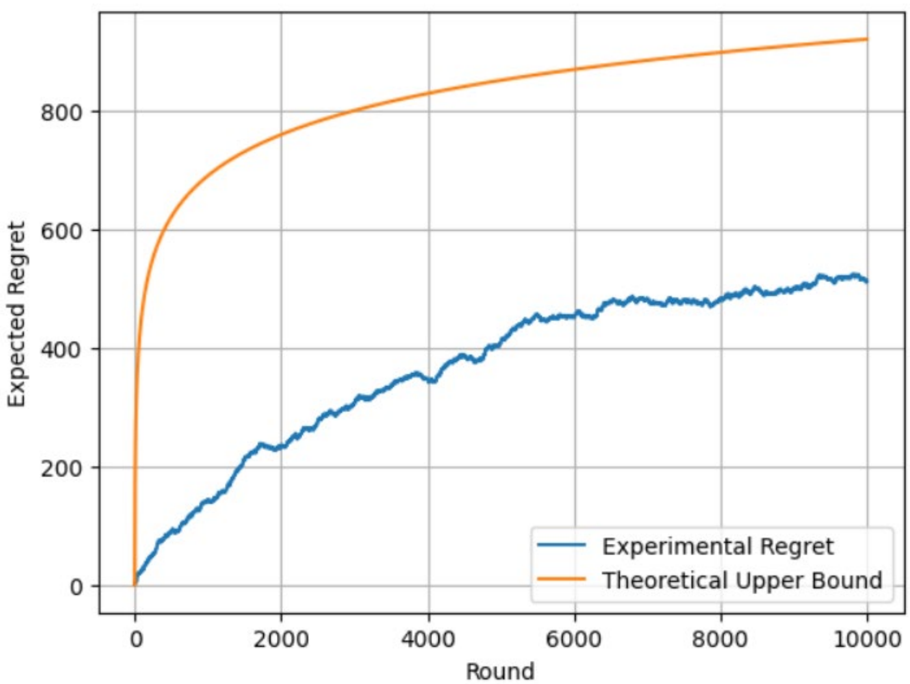
Assignment 1
Georgios Gialouris
ID: 2019030063





In the first plot we observe the regret per iteration gradually decreasing and tending towards 0.15 as we reach 1000 rounds. The upper bound for the theoretical expected regret of the modified UCB algorithm turned out to be O(log T), which means that there exists a constant c such that for sufficiently large T, the theoretical expected regret is bounded above by c log T. This is confirmed in the second plot, where we compare the experimental regret to 100 log T.





In the third plot we observe the regret per iteration gradually decreasing and tending towards 0.05 as we reach 10000 rounds. Additionally, due to the increase in the number of rounds, it becomes even more apparent in the last plot that the modified UCB algorithm achieves sublinear regret with an upper bound of O(log T).

**

So, given that $\Delta_{i,u} \gg 0, \forall i, u$, we can say that for the modified UCB algorithm

$$E[R(T)] = O(\log T)$$

Proof of the upper bound

In the original UCB algorithm the formula is:

$$ucb_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{2\log T}{N_i(t)}}$$

and the expected regret is: $E[R(T)] = \sum_{i=1}^{k} N_i(T) \cdot \Delta_i$

For the modified UCB algorithm we need to take into account the different types of users. So, the formula becomes:

$$ucb_{i,u}(t) = \hat{\mu}_{i,u}(t) + \sqrt{\frac{2\log T}{N_{i,u}(t)}}$$

, where $\hat{\mu}_{i,u}(t)$: empirical average reward of article i for user type u before round t

$N_{i,u}(t)$: number of times article i has been shown to user type u before round t

and the expected regret is: $E[R(T)] = \sum_{u=1}^{U} \sum_{i=1}^{k} N_{i,u}(T) \cdot \Delta_{i,u}$

$\Delta_{i,u} = \mu_u^* - \mu_{i,u}$, where μ_u^*: mean reward of best article for user type u

$\mu_{i,u}$: mean reward of article i for user type u

From Hoeffding's Inequality we can derive that:

Good Event: $P(Good) = P\left(\forall i,u,t : |\hat{\mu}_{i,u}(t) - \mu_{i,u}| \leq \sqrt{\frac{2\log T}{N_{i,u}(t)}}\right) = 1 - P(Bad)$

Bad Event: $P(Bad) = P\left(\exists i,u,t : |\hat{\mu}_{i,u}(t) - \mu_{i,u}| > \sqrt{\frac{2\log T}{N_{i,u}(t)}}\right) \leq k \cdot T \cdot T^{-4} = k \cdot T^{-3}$

Assume article i was chosen for user type u at round t:

$\hat{\mu}_{i,u}(t) - \mu_{i,u} \leq \sqrt{\frac{2\log T}{N_{i,u}(t)}} \Leftrightarrow \mu_{i,u} + \sqrt{\frac{2\log T}{N_{i,u}(t)}} \geq \hat{\mu}_{i,u}(t) \Leftrightarrow \mu_{i,u} + 2\sqrt{\frac{2\log T}{N_{i,u}(t)}} \geq \hat{\mu}_{i,u}(t) + \sqrt{\frac{2\log T}{N_{i,u}(t)}}$

$\geq \hat{\mu}_u^* + \sqrt{\frac{2\log T}{N_u^*(t)}}$ (since $ucb_{i,u} \geq ucb_u^*$) $\geq \mu_u^*$ (since optimal arm is also in confidence interval)

$\Rightarrow \mu_u^* - \mu_{i,u} \leq 2\sqrt{\frac{2\log T}{N_{i,u}(t)}} \Leftrightarrow \Delta_{i,u} \leq 2\sqrt{\frac{2\log T}{N_{i,u}(t)}} \Rightarrow N_{i,u}(t) \leq \frac{8\log T}{\Delta_{i,u}^2}$

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

$N_{i,u}(T) \cdot \Delta_{i,u} \leq T$ (In the bad event, we might play a terrible arm for the entire duration, hence earning max regret)

So, $P(Bad) \cdot \sum_{u=1}^{U}\sum_{i=1}^{k} N_{i,u}(T) \cdot \Delta_{i,u} \leq k \cdot T^{-3} \cdot T = k \cdot T^{-2} \to 0$ (as T grows), so we can ignore

$E[R(T)] \leq \sum_{u=1}^{U}\sum_{i=1}^{k} N_{i,u}(T) \cdot \Delta_{i,u}$ (since $P(Good) \to 1$)

$E[R(T)] \leq \sum_{u=1}^{U}\sum_{i=1}^{k} \frac{8\log T}{\Delta_{i,u}}$ (since $N_{i,u}(t) \leq \frac{8\log T}{\Delta_{i,u}^2}$ in the good event) **