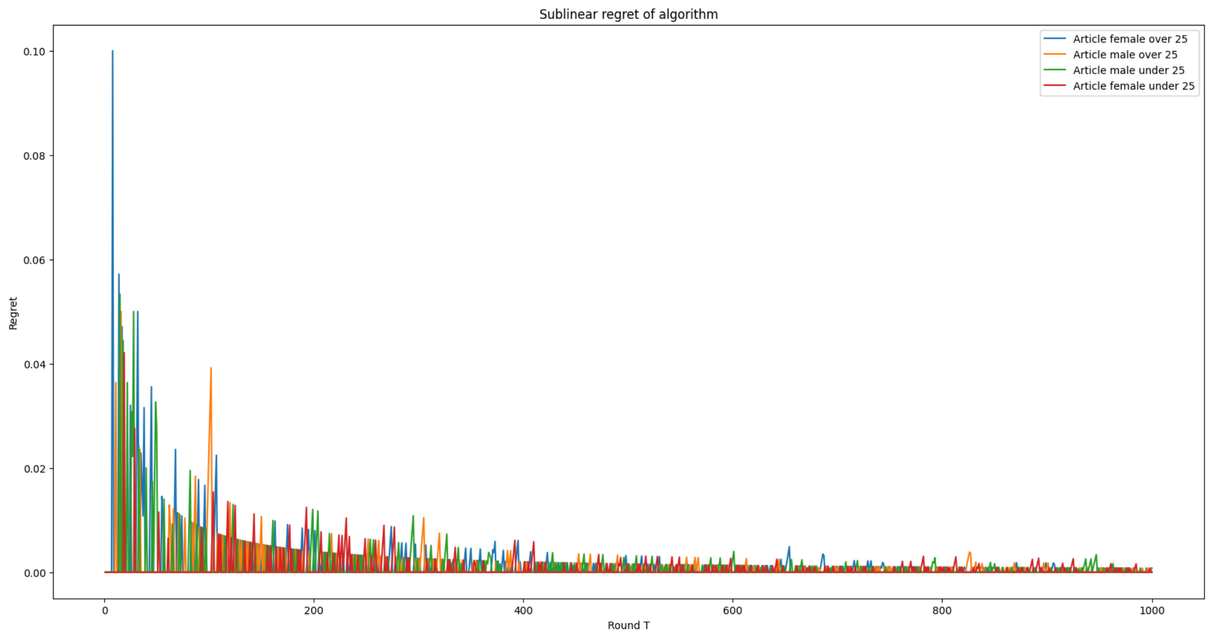
Reinforcement Learning and Dynamic Optimization (ΠΛΗ423/ΠΛΗ723)

Assignment 1

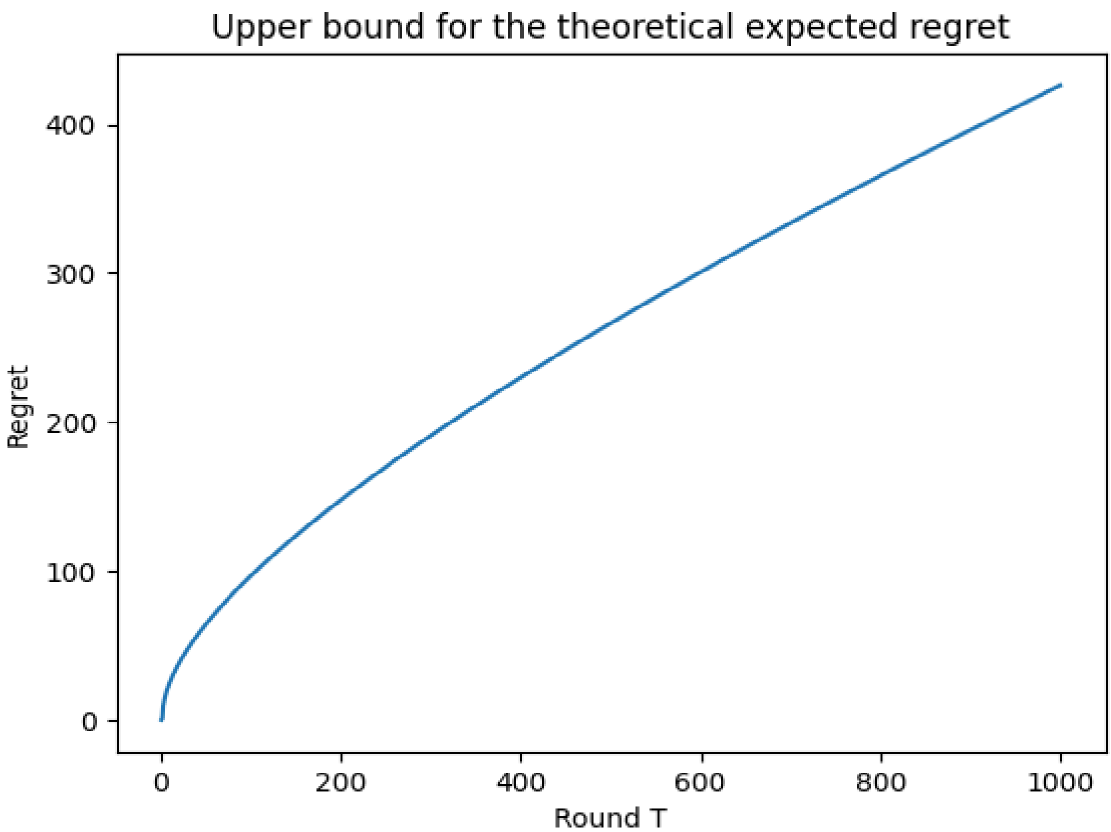
Recommending News Articles to Unknown Users

ΠΟΛΥΤΕΧΝΕΙΟ
ΚΡΗΤΗΣ

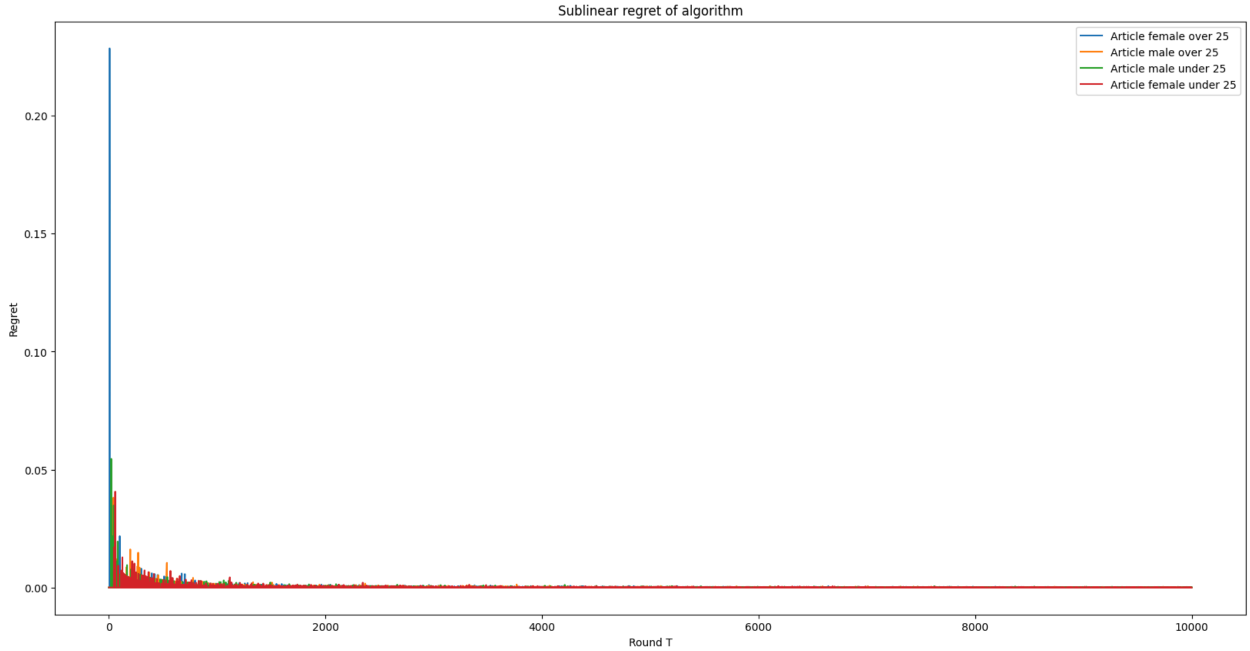
Νικόλαος Αγγελίδης 2019030190



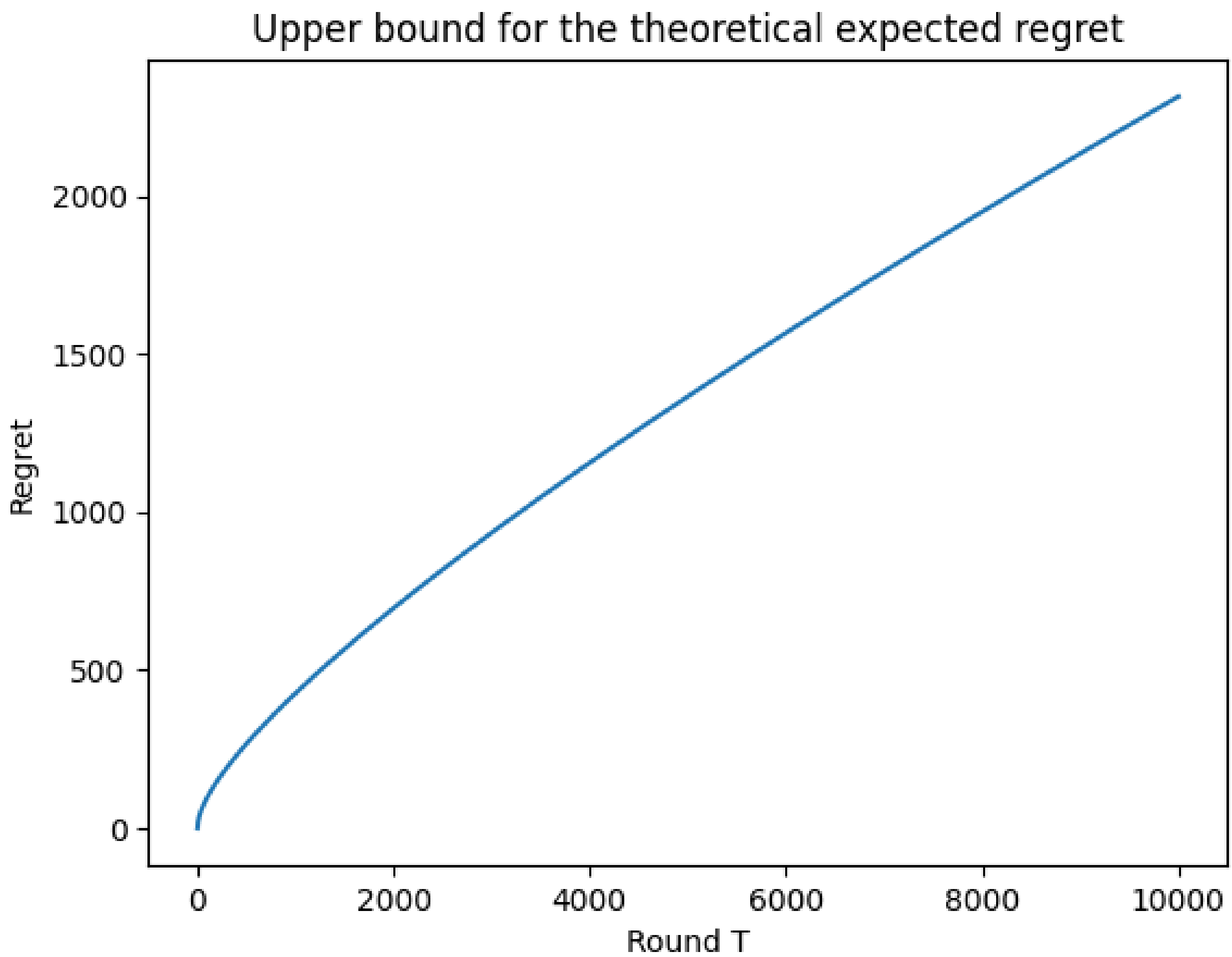
As shown in the measurement plot the regret goes to zero



Upper bound for the theoretical expected regret of the algorithm (T = 1000)



Same as before the regret inclines to zero, much faster than before.



Upper bound for the theoretical expected regret of the algorithm (T = 10000)

$$E[R(T)] \triangleq E[Regret(T)]$$

$$E[R(T)] = E[R(T)|explore] + E[R(T)|exploit]$$

$$\cdot\ E[R(T)|explore] = \sum_{t=1}^{NKU} \mu^* - \mu_{i_t} = \sum_{i=1}^{K} N \cdot U \cdot \Delta_i \leqslant N \cdot U \cdot K\, (\Delta_i \leqslant 1)$$

$$\cdot\ E[R(T)|exploit] = (T - NUK) \cdot E[\Delta_{iALG}]$$

$$E[R(T)|exploit] \leqslant T \cdot E[\Delta_{iALG}]\ (since\ T - NUK \leqslant T)$$

$$\leqslant T \cdot (1 - P(Bad))\Delta_{iALG|Good} + T \cdot P(Bad)$$

$$E[R(T)] \leq N \cdot K \cdot U + T \cdot E[\Delta_{iALG}]$$

$$* \epsilon = \sqrt{2\, log(T) / NU}$$

$$* P(Bad) = P\left(\exists i : |\mu\hat{}i - \mu i| > \sqrt{2\, log(T) / NU}\right) \leqslant KT^{-4}$$

$$* P(Good) = P\left(\forall i : |\mu\hat{}i - \mu i| \leqslant \sqrt{2\, log(T) / NU}\right) \geqslant 1 - KT^{-4}$$

$$* \Delta_{iALG} = \mu * - \mu i \leq \mu * - \mu\hat{}i + \epsilon\ (since\ \mu i \geq \mu\hat{}i - \epsilon)$$

$$\leq \mu\hat{} * + \epsilon - \mu\hat{}i + \epsilon (since\ \mu * \leq \mu\hat{}i + \epsilon) \Rightarrow \Delta_{iALG|Good} \leq 2\epsilon$$

$$E[R(T)] \leq N \cdot K \cdot U + T \cdot (1 - P(Bad)) \cdot \Delta_{iALG|Good} + T \cdot P(Bad)$$

$$E[R(T)] \leq N \cdot K \cdot U + T \cdot \left(1 - KT^{-4}\right) \cdot \Delta_{iALG|Good} + T \cdot KT^{-4}$$

$$E[R(T)] \leq N \cdot K \cdot U + T \cdot \Delta_{iALG|Good}\ \left(since\ KT^{-3} \to 0,\ as\ T\ grows\right)$$

$$E[R(T)] \leq N \cdot K \cdot U + T \cdot 2\sqrt{2\, log(T) / NU}$$

$$We\ want:\ N \cdot K \cdot U \approx T \cdot 2\sqrt{2\, log(T) / NU}$$

$$N = O\left(\left(\frac{TK^2 log(T)}{U^2}\right)^{1/3}\right)\ (1)$$

$Insert\ (1)\ into\ the\ regret\ formula:$

$$E[R(T)] \leq O\left(\left(\frac{TK^2 log(T)}{U^2}\right)^{1/3}\right) \cdot K \cdot U + T \cdot 2\sqrt{\frac{2\, log(T)}{O\left(\left(\frac{TK^2 log(T)}{U^2}\right)^{1/3}\right)U}}$$

$*\ First\ term:$

$$O\left(\left(\frac{TK^2 log(T)}{U^2}\right)^{1/3}\right) \cdot K \cdot U = O(\left(T^{1/3} K^{2/3} log(T)^{1/3}\right) \cdot K / U^{1/3} =$$

$$O\left(\frac{T^{1/3} K^{5/3} log(T)^{1/3}}{U^{1/3}}\right)$$

$*\ Second\ term:$

$$T \cdot 2\sqrt{\frac{2\, log(T)}{O\left(\left(\frac{TK^2 log(T)}{U^2}\right)^{1/3}\right)U}} = O\left(T \cdot \sqrt{\frac{log(T)}{T^{1/3} K^{2/3} log(T)^{1/3} / U^{2/3}}}\right) = O\left(T \cdot \sqrt{\frac{U^{2/3}}{T^{1/3} K^{2/3}}}\right)$$

$So\ the\ theoretical\ expected\ regret\ of\ the\ algorithm:$

$$E[R(T)] \leq O\left(\frac{T^{1/3} K^{5/3} log(T)^{1/3}}{U^{1/3}}\right) + O\left(T \cdot \sqrt{\frac{U^{2/3}}{T^{1/3} K^{2/3}}}\right)$$