

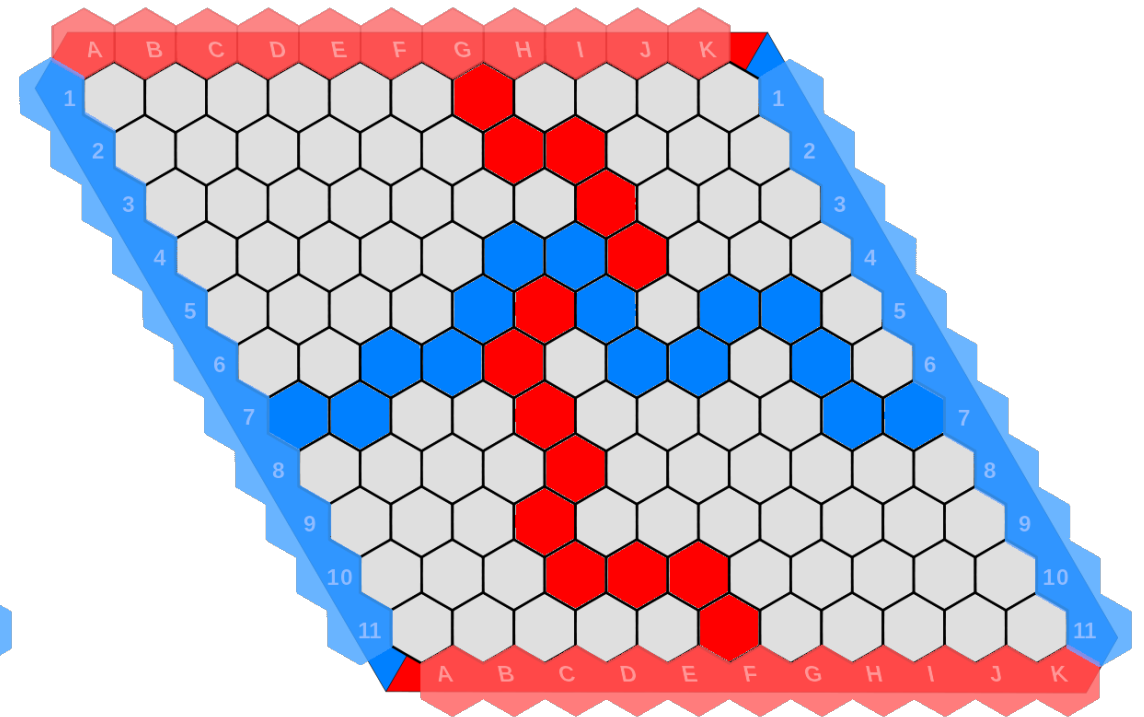
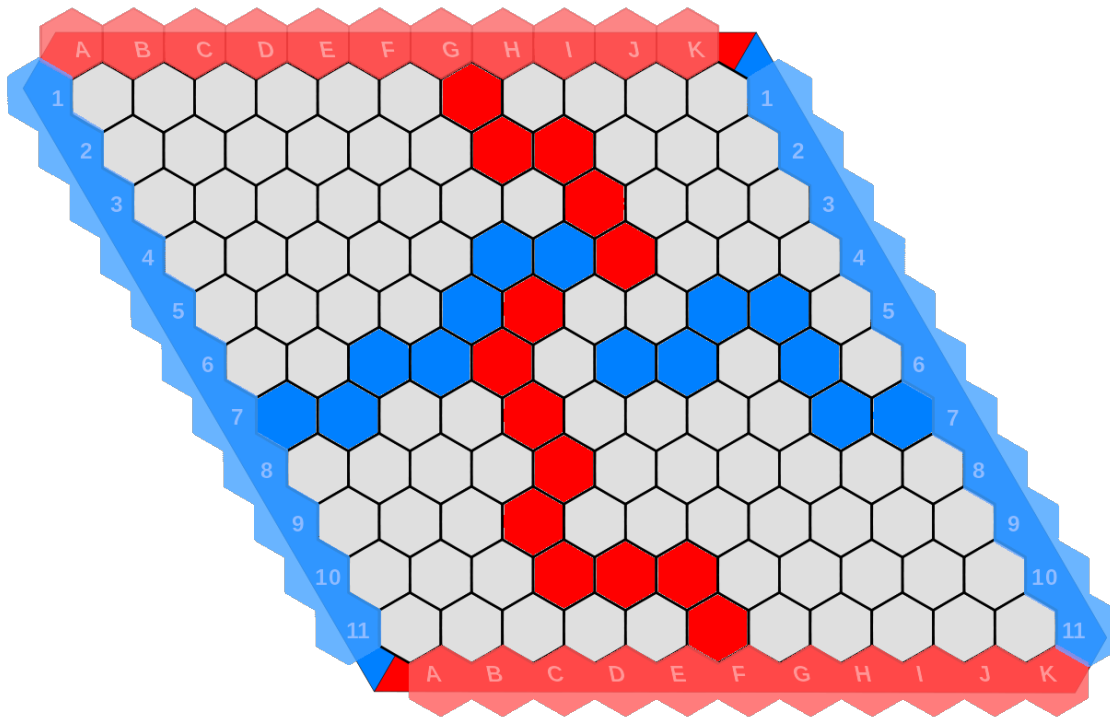


DES CASINOS À L'INTELLIGENCE ARTIFICIELLE

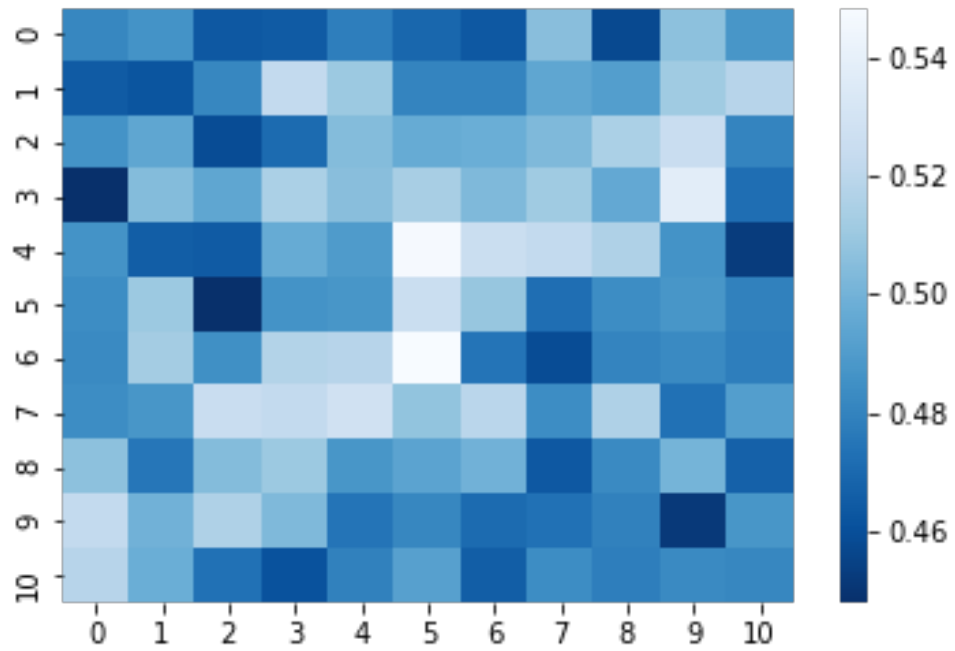
Travail Encadré de Recherche
M1 DS

Le jeu de Hex

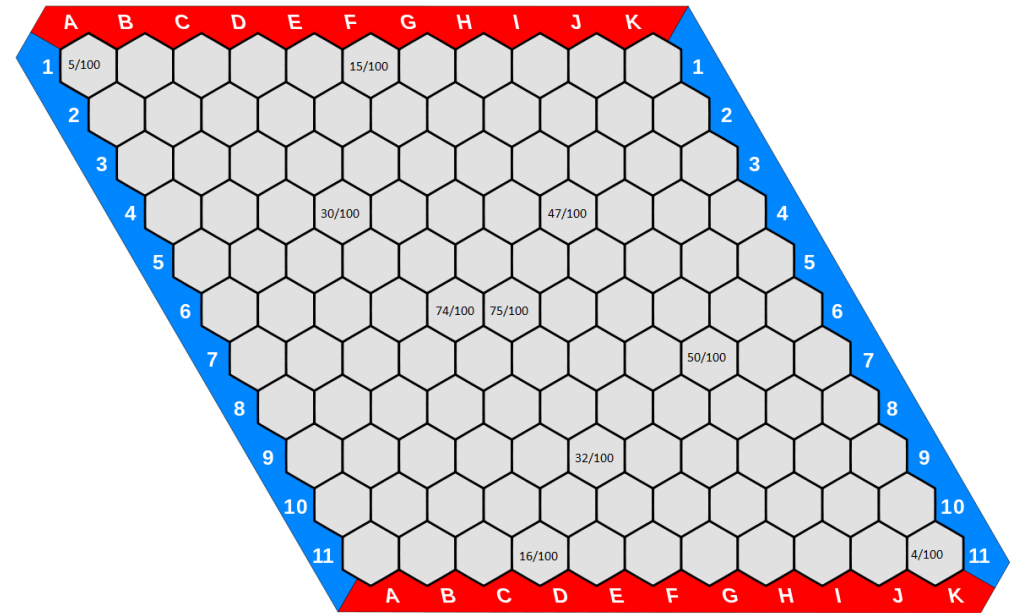
Condition de fin de partie



Monte-Carlo



Probabilité de gagner la partie



UCB et problème du bandit manchot



Theorem 1. For all $K > 1$, if policy UCB1 is run on K machines having arbitrary reward distributions P_1, \dots, P_K with support in $[0, 1]$, then its expected regret after any number n of plays is at most

$$\left[8 \sum_{i: \mu_i < \mu^*} \left(\frac{\ln n}{\Delta_i} \right) \right] + \left(1 + \frac{\pi^2}{3} \right) \left(\sum_{j=1}^K \Delta_j \right)$$

where μ_1, \dots, μ_K are the expected values of P_1, \dots, P_K .

Deterministic policy: UCB1.

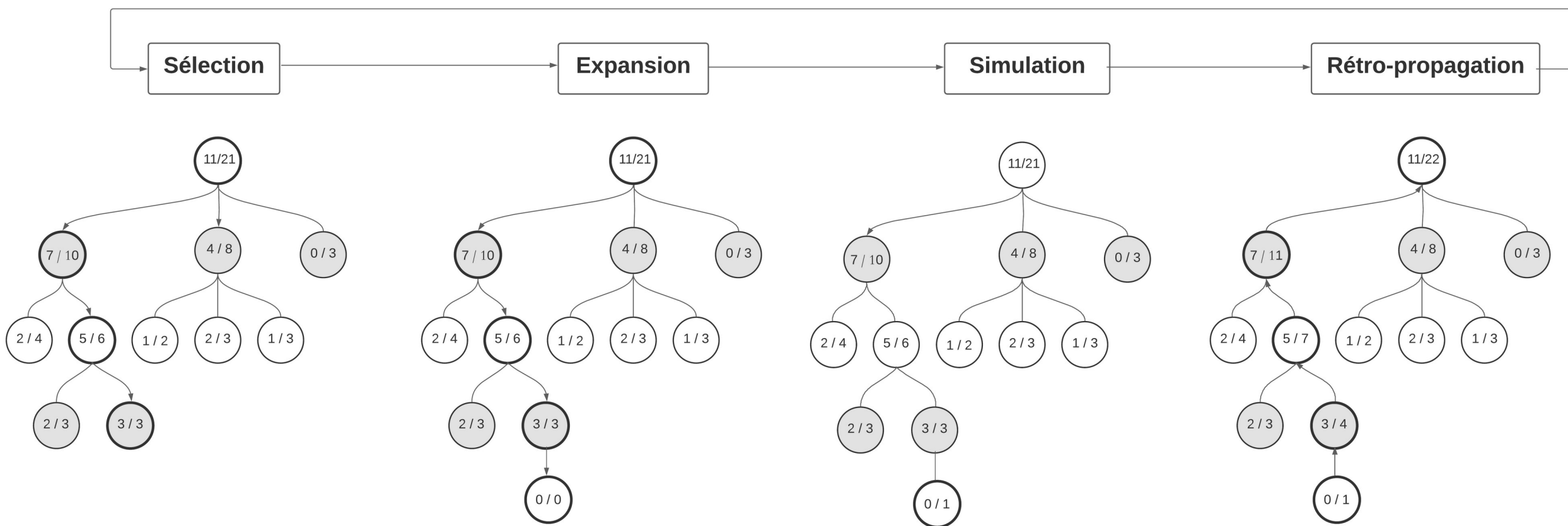
Initialization: Play each machine once.

Loop:

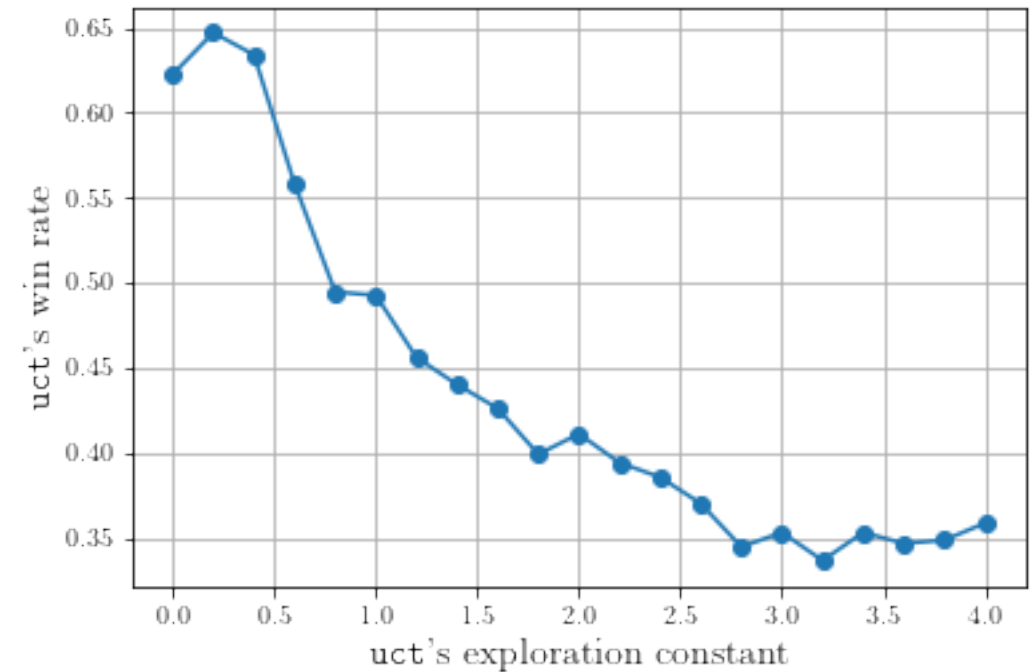
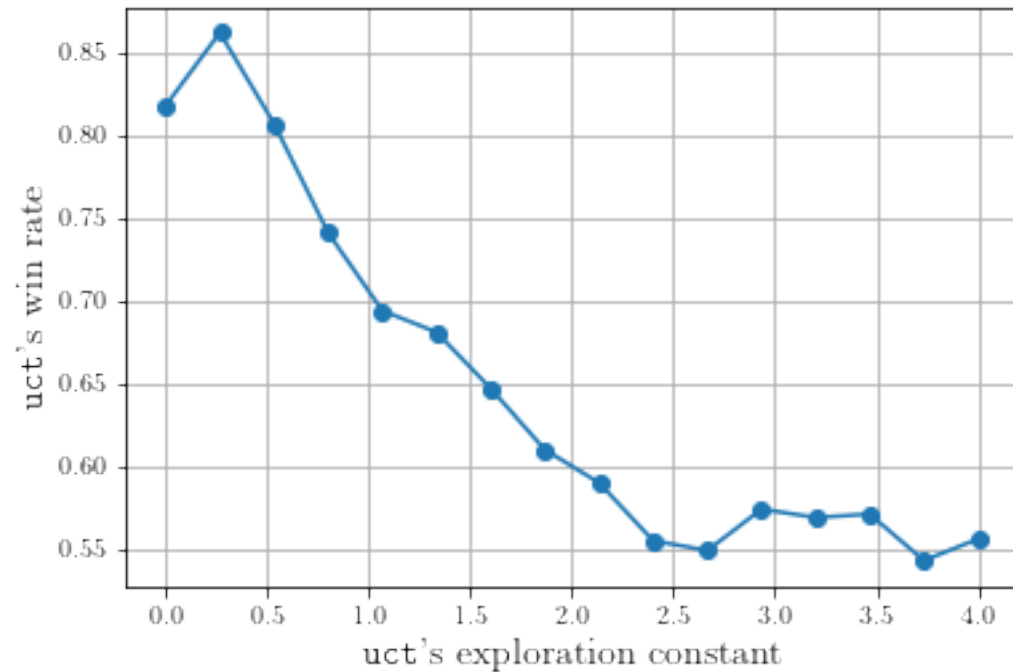
- Play machine j that maximizes $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$, where \bar{x}_j is the average reward obtained from machine j , n_j is the number of times machine j has been played so far, and n is the overall number of plays done so far.

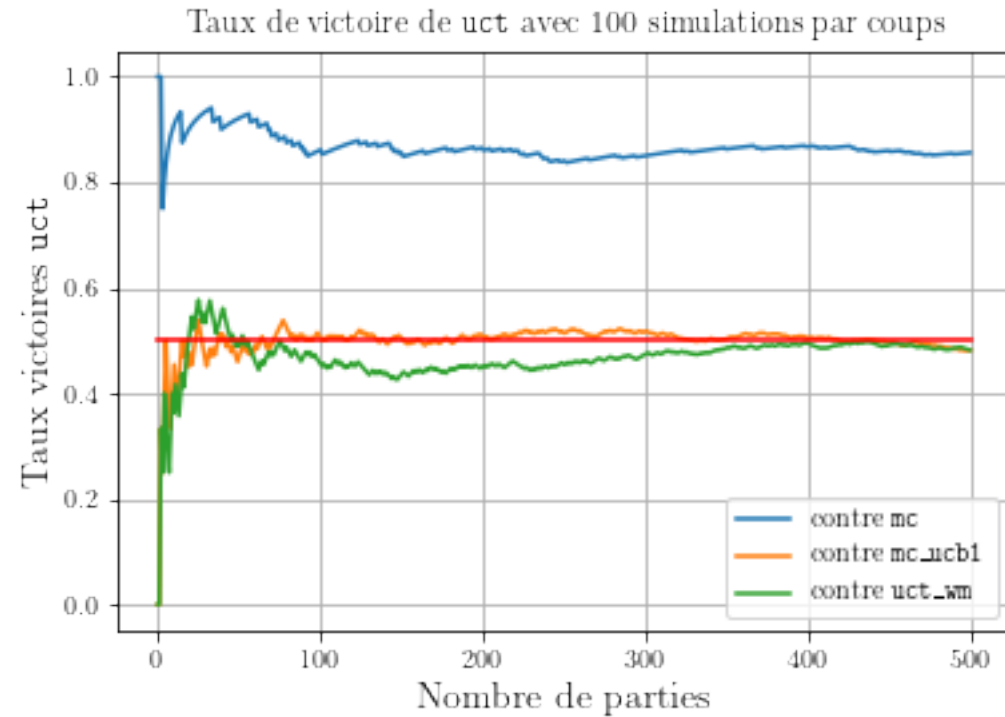
Figure 1. Sketch of the deterministic policy UCB1 (see Theorem 1).

Upper Confidence Bound applied to Trees (UCT)

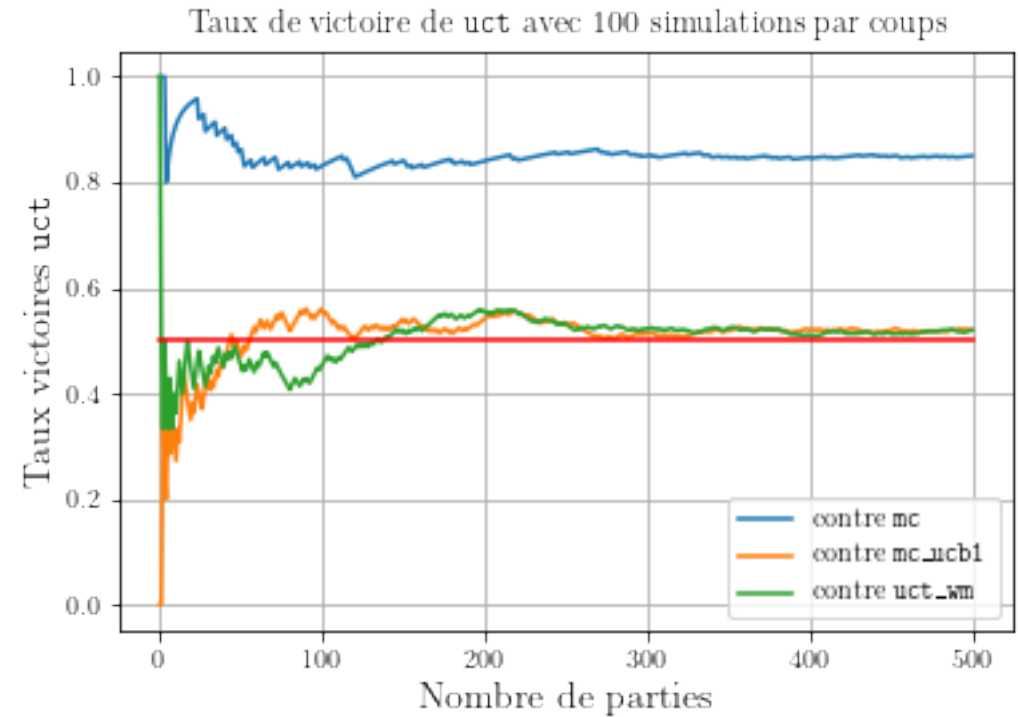


Constante d'exploration optimale

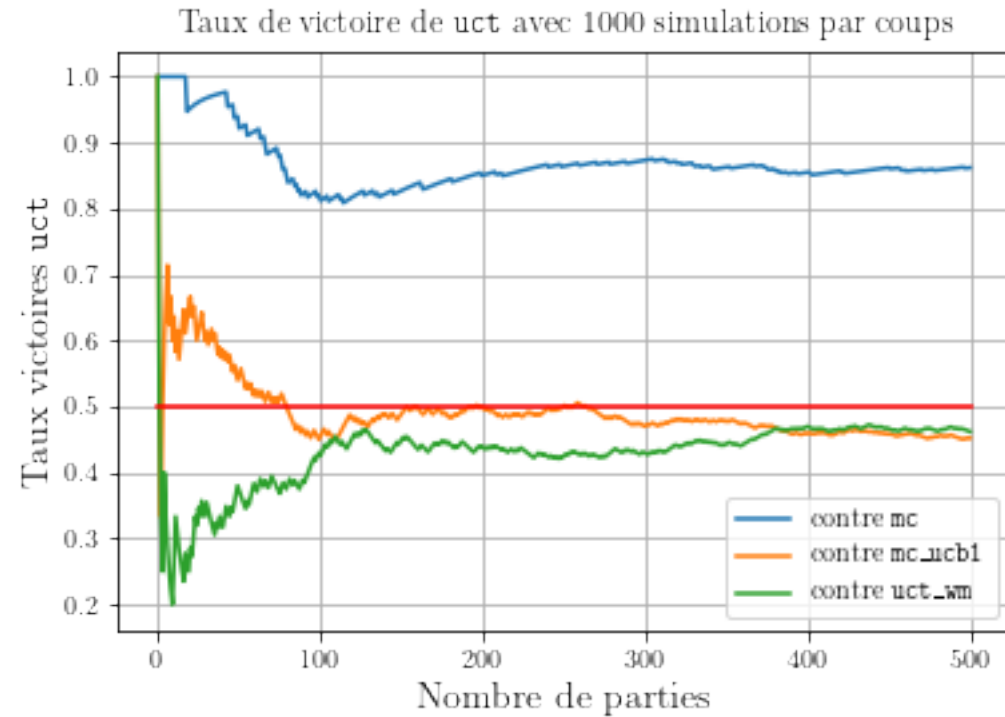




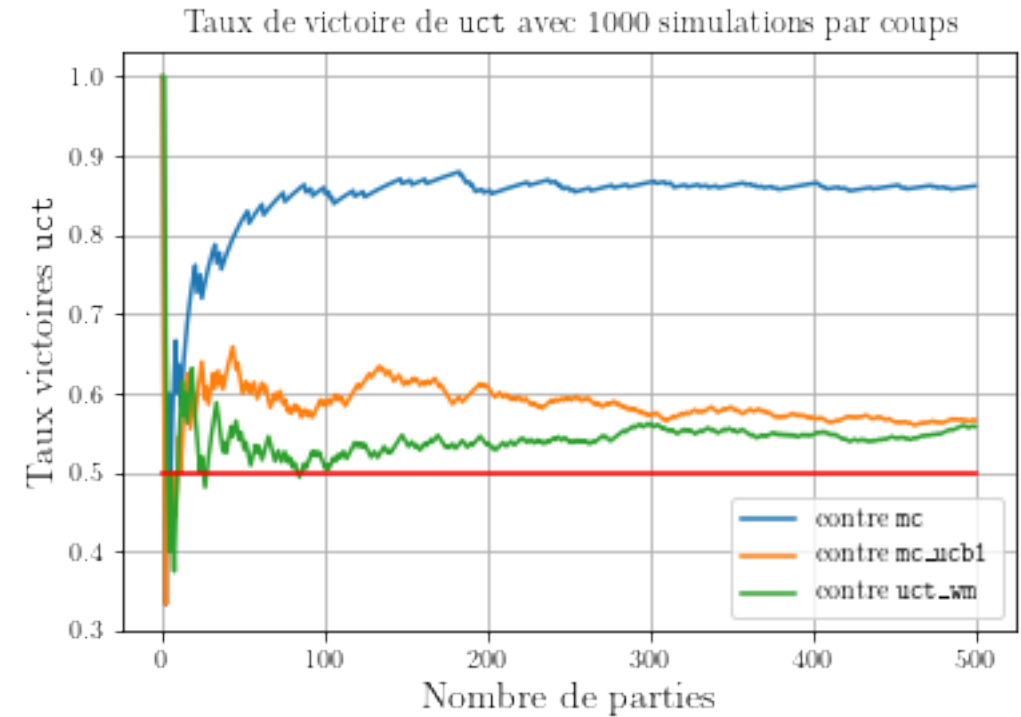
(a) uct joue en premier



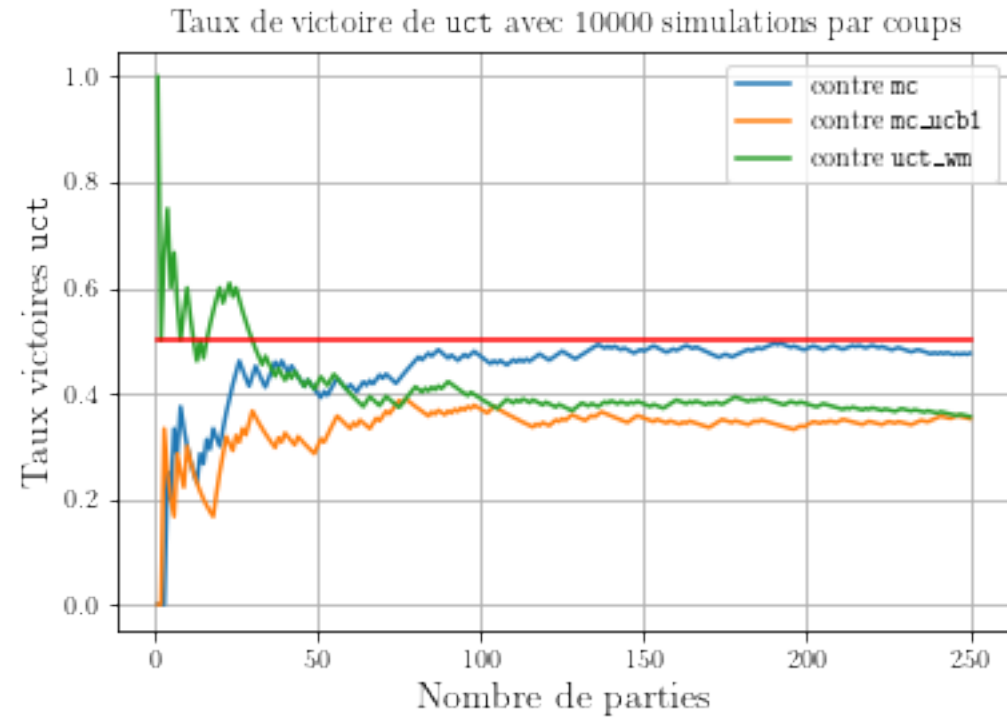
(b) uct joue en deuxième



(a) uct joue en premier



(b) uct joue en deuxième



(a) uct joue en premier



(b) uct joue en deuxième

Merci pour votre attention

Avez-vous des questions ?