

Apprentissage par renforcement et Méthode de Monte Carlo

Igor Stéphan

UFR Sciences Angers

2018-2019

- « Supervisé » (un ensemble d'exemples et de contre-exemples)
- « Non supervisé » (pour découvrir des régularités)
- Algorithme d'apprentissage (de réflexes) par renforcement

Apprentissage (de réflexes) par renforcement

- Apprentissage de réflexes (sans réflexion)
- Récompenses (/punitions) : signal de renforcement
- Connaissance possiblement imparfaite (stochastique) de l'environnement
 - des états (et/ou)
 - des transitions (état, action, successeur dans le graphe des états)
- Connaissance imparfaite (voire nulle) de l'environnement sur les signaux de renforcement des états (et/ou) des transitions
- Perception $\rightsquigarrow$ action $\rightsquigarrow$ récompense
- Connaissance locale de la topologie (de fait ou due à la complexité)
- Hypothèse : « les règles du jeu » sont stationnaires au cours du temps
- Système très lent mais sous hypothèse faible : très large champs d'application

- Apprentissage d'une fonction des couples (état,action) dans $\mathbb{R}$: une politique qui calcule une probabilité
- Apprentissage d'une fonction V d'espérance de gain des états
- Apprentissage d'une fonction Q d'espérance de gain des couples (état/action)
- Les mesures de gain
 - Cumul à horizon finie : $\sum_{t=0}^{T-1} r_t$
 - Cumul avec intérêt γ : $\sum_{t=0}^{\infty} \gamma^t r_t$
 - Gain en moyenne : $\frac{1}{T-1} \sum_{t=0}^{T-1} r_t$
 - *In fine* (victoire/défaite/partie nulle)
- Mesure de l'efficacité de la méthode d'apprentissage : convergence plus ou moins rapide vers la politique optimale
- Parcours d'un espace de recherche : exploration vs exploitation
 - exploration : explorer de toutes nouvelles possibilités
 - exploitation : intensifier dans les zones montrées comme prometteuses

Le Monte Carlo Tree Search (MCTS)

- Echantillonnage de scenarios
- Apprentissage *self-play*
- Construction de l'arbre de recherche
 - renforce ses connaissances
 - accroît l'étendue de ses connaissances
- Méthode *any-time*
- Principe min-Max

- Initialisation : état initial, seul nœud de l'arbre de recherche
- Pour k scénarios simulés :
 - ① La descente (*descent*)
 - Selection des coups sur l'arbre de recherche actuel
 - Fonction de valeur de coup : $Q(e : \text{état}, c : \text{coup})$
 - Choix (i.e. choix du nœud renforcé) d'un candidat non exploré puis
 - choix du meilleur candidat
 - Compromis exploitation vs exploration
 - ② L'accroissement (*growth*)
 - Sur une feuille de l'arbre de recherche actuel
 - Choix d'un nouveau nœud à créer parmi les successeurs de l'état étiquetant la feuille
 - ③ L'extension (*roll-out*)
 - Poursuite du reste du scénario de manière aléatoire
 - ④ La mise-à-jour (*update*)
 - Rétro-propagation de la valeur du scénario

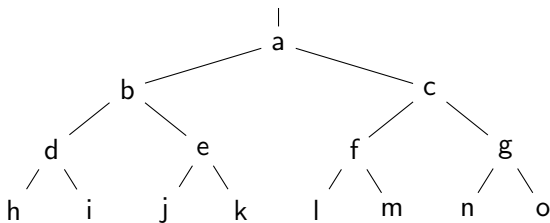
- Aléatoire avec une pondération selon la rétro-propagation
- Déterministe
 - Upper Bound Confidence (UBC) :

$$Q_{UBC}(e, c) = \underbrace{\overline{Q(e, c)}}_{\text{exploitation}} + \underbrace{\sqrt{C \times \frac{\ln k}{n(e, c)}}}_{\text{exploration}}$$

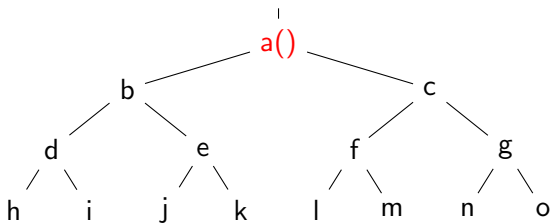
- $\overline{Q(e, c)}$: gain moyen du coup c sur l'état e
- C : constante de pondération exploration vs exploitation
- k : nombre d'itérations à partir de ce nœud
- $n(e, c)$: nombre de sélection du coup c sur l'état e (localement)

UCT (*Upper Bound confidence Tree*) = MCTS + UBC

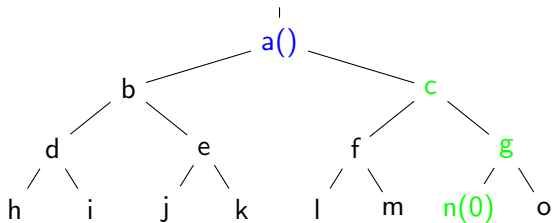
Exemple (avec $C = 2$)



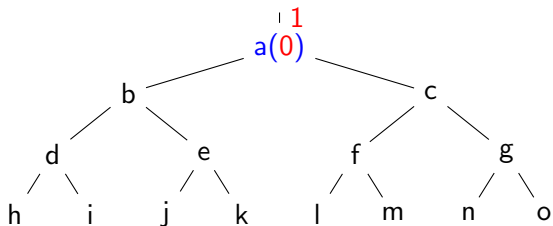
1 : pas de *descente*; *growth*



1 : *roll out*



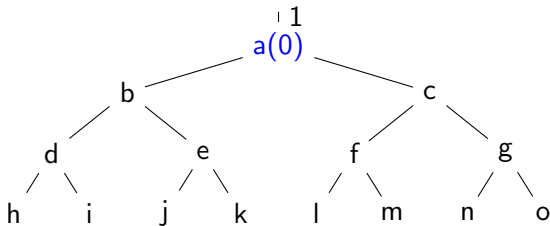
1 : *update*



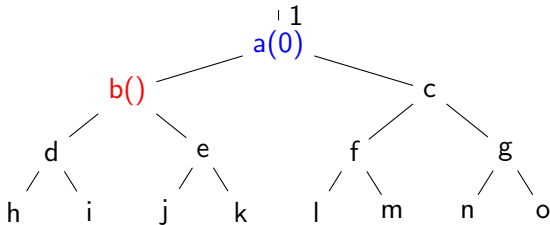
UCT (*Upper Bound confidence Tree*) = MCTS + UBC

Exemple (avec $C = 2$)

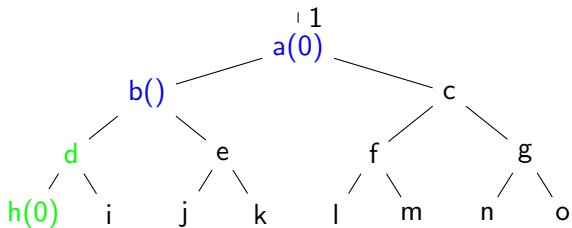
2 : *descente*



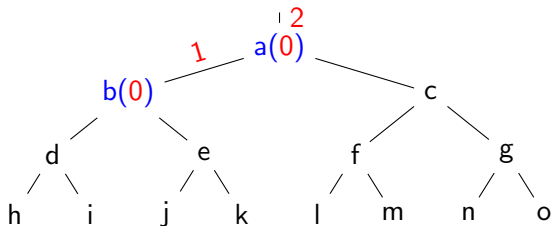
2 : *growth*



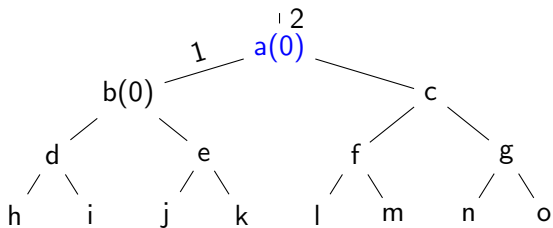
2 : roll out



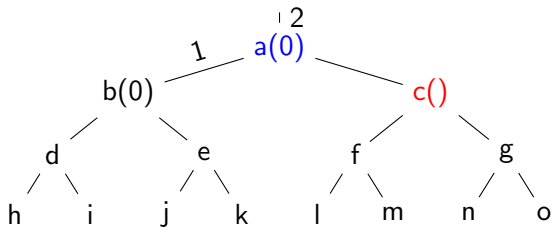
2 : update



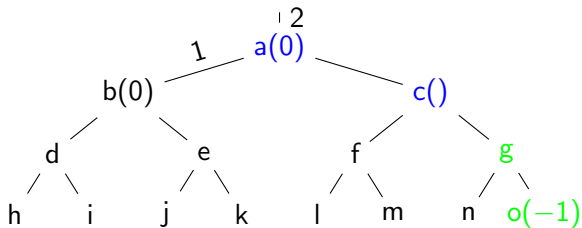
3 : *descente*



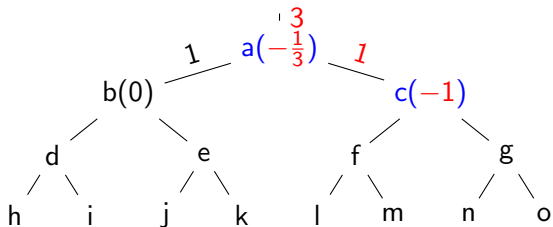
3 : *growth*



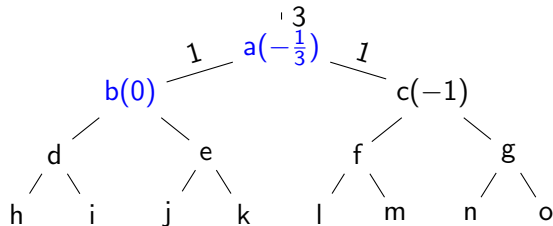
3 : roll out



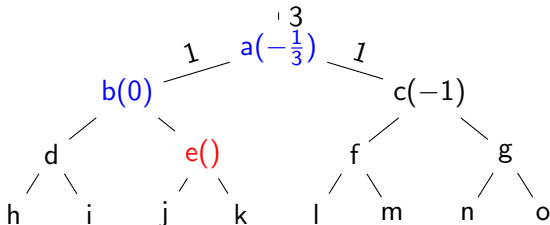
3 : update



4 : *descente*



4 : *growth*



$$Q_{UCB}(a, gauche)$$

$$= 0 + \sqrt{2 \times \frac{\ln 3}{1}}$$

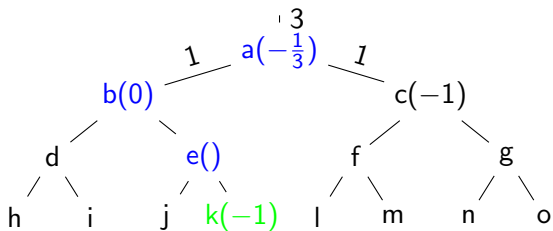
$$\approx 1.482$$

$$Q_{UCB}(a, droite)$$

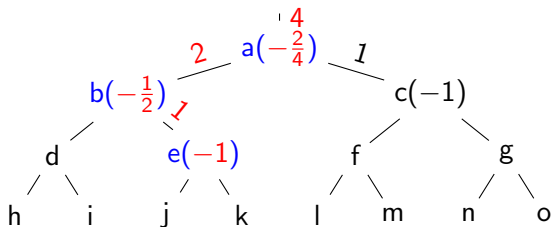
$$= -1 + \sqrt{2 \times \frac{\ln 3}{1}}$$

$$\approx 0.482$$

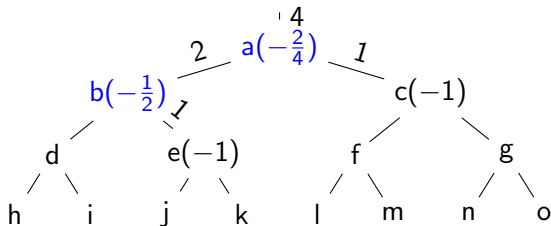
4 : *roll out*



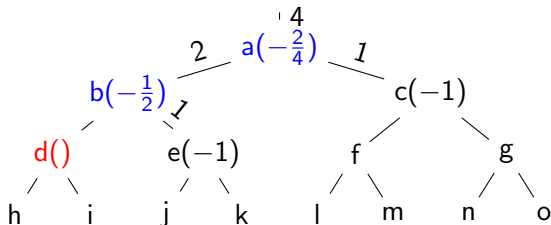
4 : *update*



5 : *descente*



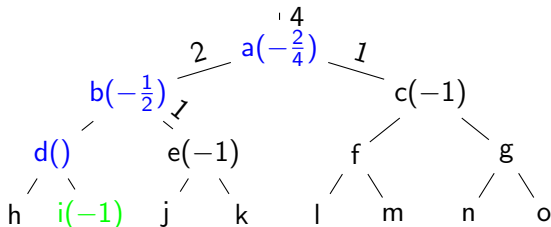
5 : *growth*



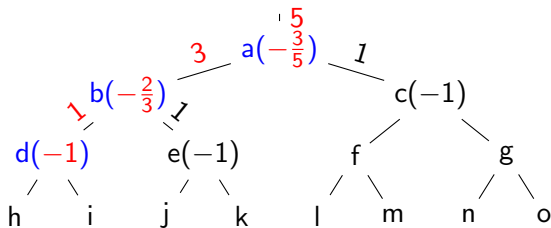
$$Q_{UCB}(a, gauche) = -0.5 + \sqrt{2 \times \frac{\ln 4}{2}} \approx 0.677$$

$$Q_{UCB}(a, droite) = -1 + \sqrt{2 \times \frac{\ln 4}{1}} \approx 0.665$$

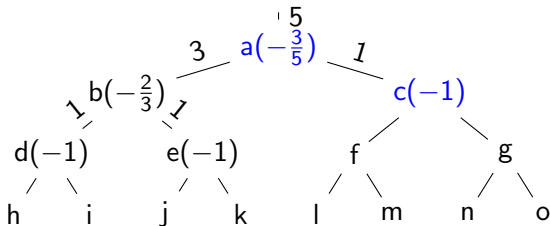
5 : *roll out* (le nœud *h* aurait pu être choisi à nouveau)



5 : *update*



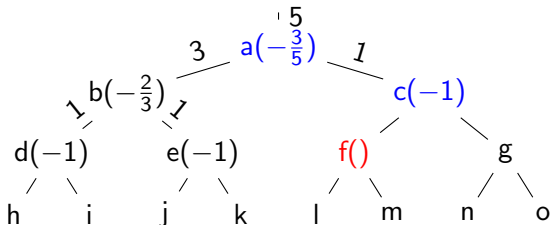
6 : *descente*



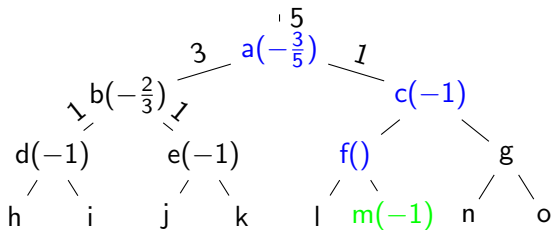
$$Q_{UCB}(a, gauche) \\ = -\frac{2}{3} + \sqrt{2 \times \frac{\ln 5}{3}} \\ \approx 0.369$$

$$Q_{UCB}(a, droite) \\ = -1 + \sqrt{2 \times \frac{\ln 5}{1}} \\ \approx 0.794$$

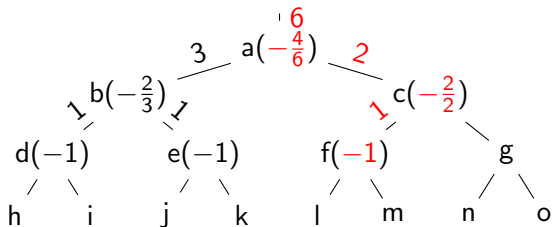
6 : *growth*



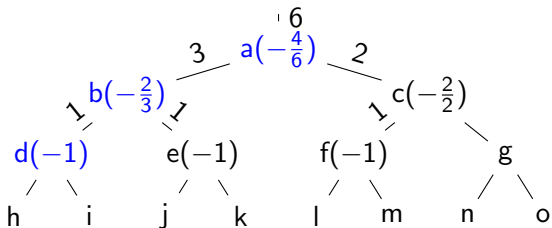
6 : *roll out*



6 : *update*



7 : *descente*



$$Q_{UCB}(a, gauche) = -\frac{2}{3} + \sqrt{2 \times \frac{\ln 6}{3}} \approx 0.426$$

$$Q_{UCB}(a, droite) = -\frac{2}{2} + \sqrt{2 \times \frac{\ln 6}{2}} \approx 0.339$$

$$Q_{UCB}(b, gauche) = -1 + \sqrt{2 \times \frac{\ln 3}{1}} \approx 0.982 = Q_{UCB}(b, droite)$$