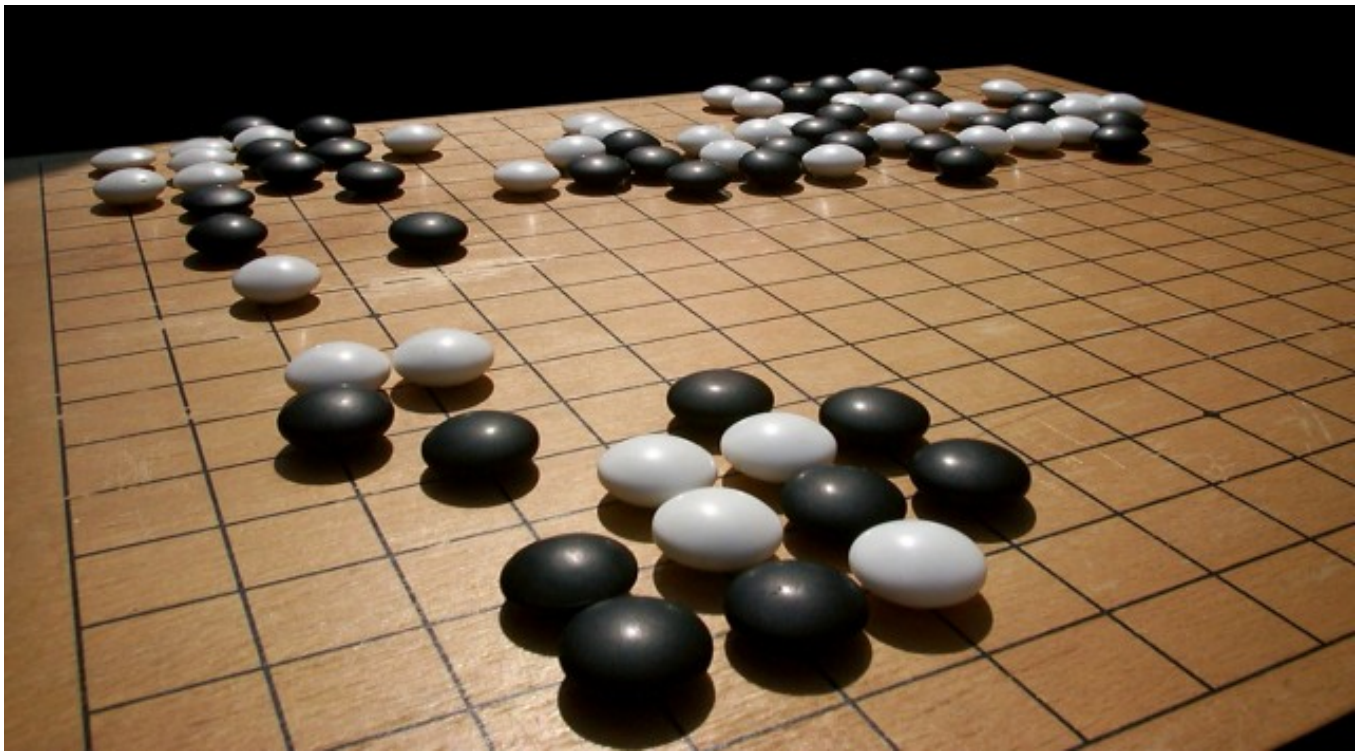
Jeu de Go et Exploration d'Arbre par Bandit

CentraleSupélec – Gif

IA et Jeu de Go

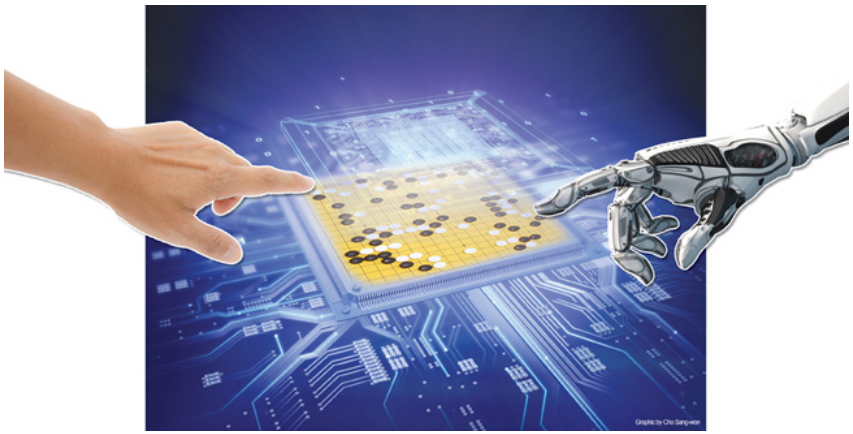


Plan

Pourquoi une IA pour un jeu ?

- Avoir une IA pour un jeu...
- Représentation des problèmes de décision
- **Environnement** bien défini : règles du jeu
- **Évaluation** facile : score
- Challenge de battre les humains

Pourquoi le jeu de Go ?



- un jeu de plateau qui a longtemps résisté aux IA
- règles **simples**
- méthodes classiques (alphabeta) inefficaces

Histoire

- aurait été inventé en chine en 2000 BC
- premiers écrits : 500 BC
- fait parti des 4 arts majeurs chinois : peinture, calligraphie, guqin, *go*
- se répand en Asie dès 800 dans la noblesse
- aujourd'hui, environ 20 millions de joueurs

Matériel

- plateau de jeu : Goban
- deux tailles 9x9 ou 19x19
- pierres noires et blanches

Règles : placement

Chaque joueur pose une pierre à tour de rôle Pierres posées sur les intersections Blanc commence

Règles : chaînes et capture

pierres reliées horizontalement ou verticalement forment une chaine
emplacement libre autour d'une chaine : liberté enlever la dernière liberté d'une chaine : capture

Règles : fin de partie

les deux joueurs passent score

Règles : le ko

illustration histoire règle humain/ordinateur

Echelle de niveau

Plan

Présentation

Alpha beta

Découpage du plateau

Règles expertes

Échelle de niveau

Plan

Introduction du problème

Dans un casino, il y a plusieurs machines à sous différentes en terme de récompense.

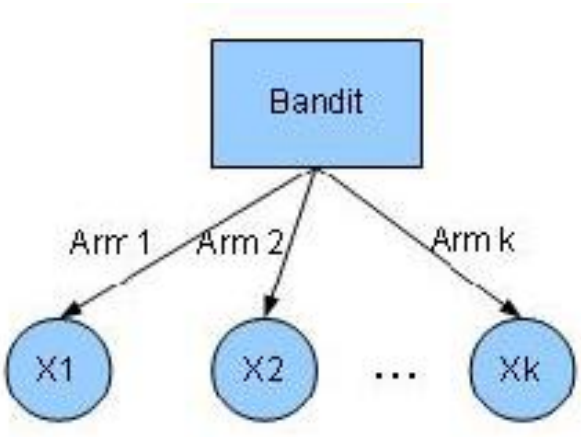
- Comment répartir mes pièces entre les machines ?

Autres problèmes similaires

- Essais cliniques : trouver le traitement qui fonctionne le mieux.
- Sélection d'un serveur dans un réseau : trouver le serveur avec le temps de réponse le plus faible.
- Publicité ciblée : trouver le type de pub qui intéressera le plus un utilisateur.
- ...

Ce sont des problèmes où on a plusieurs fois le même choix à effectuer. Le choix conduit à une récompense aléatoire.

Définition formelle



- un ensemble de bras $A = \{1, ..., K\}$.
- chaque bras est associé à une distribution de probabilité X_k d'espérance μ_k.
- l'algorithme choisit un bras a à chaque pas de temps.
- le bandit retourne une récompense r : une réalisation de X_a.
- les tirages successifs sur un même bras sont indépendant et identiquement distribués.

Notations supplémentaires

- $T_i(n)$: le nombre de fois que le bras i a été sélectionné au pas de temps n.
- $\mu^* = \max_{1 \leq i \leq K} \mu_i$
- $\Delta_i = \mu^* - \mu_i$
- $\Delta = \min_{i:\Delta_i > 0} \Delta_i$

Objectif

Le but est d'optimiser le regret R_n défini comme suit :

$$R_n = \mu^* n - \mathbb{E} \sum_{j=1}^{K} T_j(n) \mu_j$$

$$R_n = \sum_{j=1}^{K} \Delta_j \mathbb{E}[T_j(n)]$$

Borne inférieure

Pour toute stratégie d'allocation et pour tout bras non optimal :

$$\mathbb{E}[T_j(n)] \geq \frac{\log n}{D(p_j||p^*)}$$

$$\text{où } D(p_j||p^*) = \int p_j \log \frac{p_j}{p^*}$$

On en déduit que le meilleur regret atteignable est en $\log(n)$.

[Lai and Robbins, 1985]

UCB

Principe de l'algorithme :

- A partir des informations disponibles au temps t, on calcule la borne de confiance supérieur (UCB) correspondant à chaque bras.
- On choisit le bras qui a la valeur UCB la plus grande.

[Auer and all, 2002]

UCB

Calcul de la valeur UCB pour le bras i au pas de temps t :

$$\hat{\mu}_{i,t-1} + \sqrt{\frac{3\log(t)}{2T_i(t-1)}}$$

où $\hat{\mu}_{i,t-1}$ correspond à la moyenne empirique du bras i.

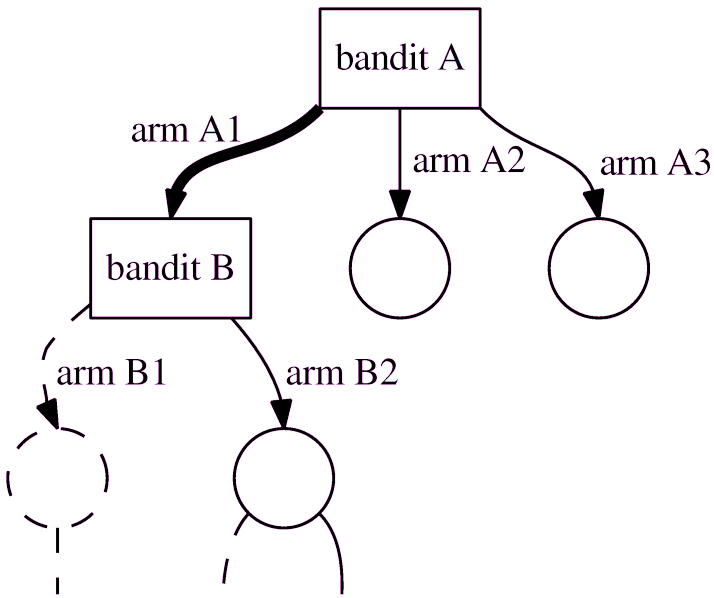
UCB

Borne sur le regret :

$$R_n \leq 6 * \sum_{i \neq i^*} \frac{\log(n)}{\Delta_i} + K(\frac{\pi^2}{3} + 1)$$

Descente dans l'arbre

La descente dans l'arbre se fait en considérant que chaque choix d'une branche est un problème de bandit.



UCB en pratique

- Ajout d'un paramètre p de contrôle de l'exploration :

$$\hat{\mu}_{i,t-1} + p\sqrt{\frac{\log(t)}{T_i(t-1)}}$$

- Ajout de connaissances a priori $C_i(t)$:

$$\hat{\mu}_{i,t-1} + p\sqrt{\frac{\log(t)}{T_i(t-1)}} + C_i(t)$$

AMAF

Ajout de Connaissances Expertes

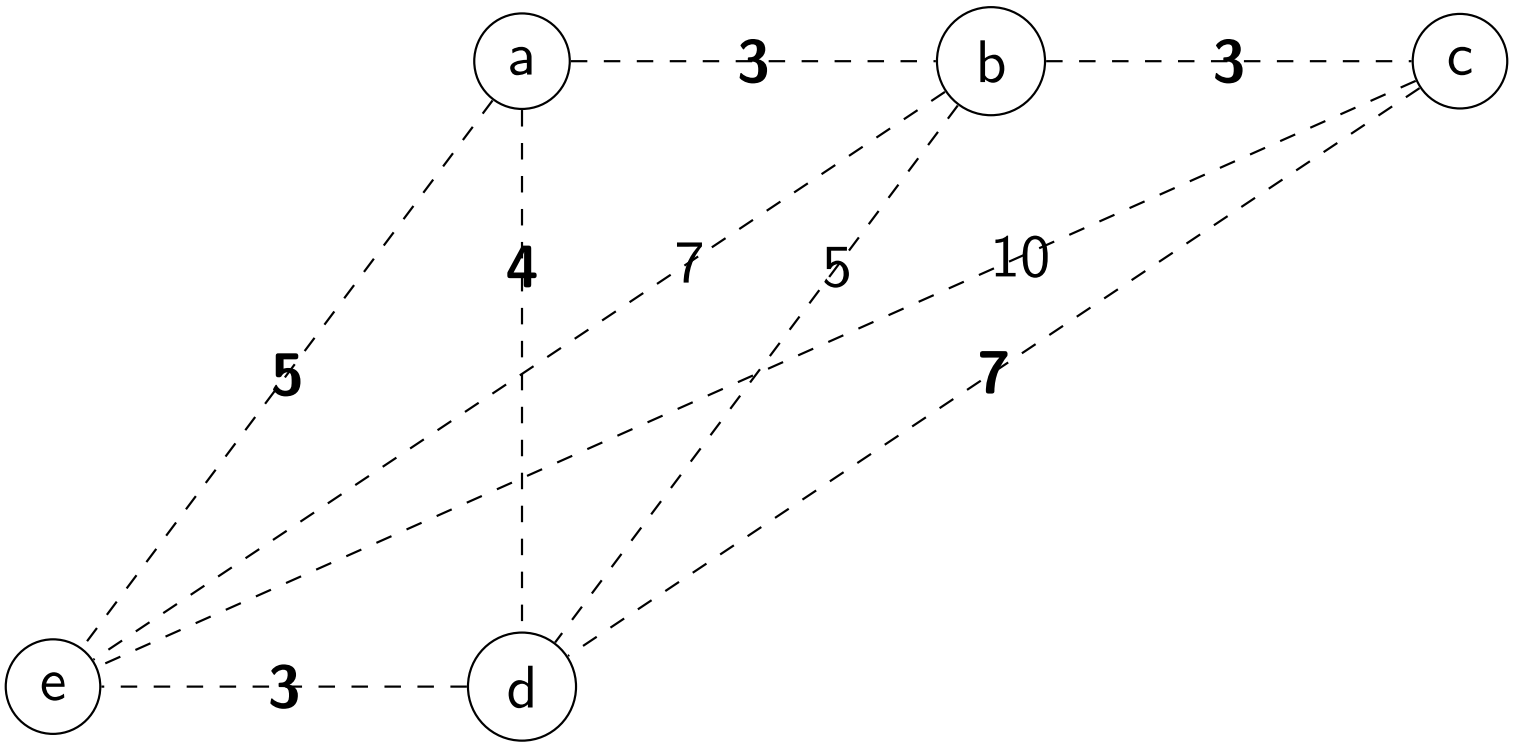
Plan

Deep Learning

Autres applications

Conclusion

Exemple borne



Calcul de la borne

trajet déjà effectué $= \emptyset$

$$(\underbrace{3+4}_{a}+\underbrace{3+3}_{b}+\underbrace{3+7}_{c}+\underbrace{3+4}_{d}+\underbrace{3+5}_{e})/2 = 19$$