Sylvain Gelly

A contribution to Reinforcement Learning: Application to Computer-Go

Inférence, Apprentissage, Optimisation

UNIVERSITÉ PARIS-SUD 11

digiteo

Research in science and technology of information

Apprentissage par Renforcement

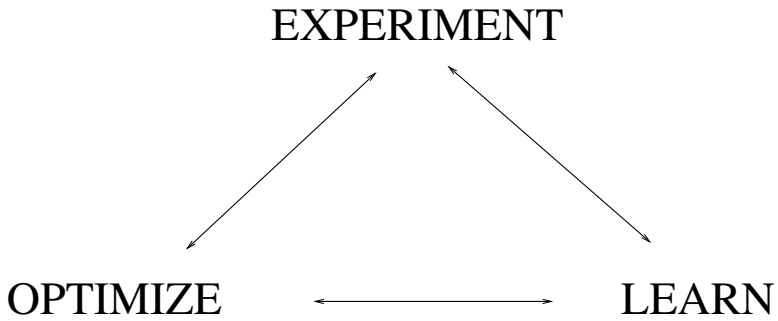
Contexte

Le monde est un labyrinthe.

Certaines actions, dans certains états, portent des fruits (*rewards*) avec un certain retard [avec une certaine probabilité].

Le but : trouver la politique (état $\longrightarrow$ action) maximisant l'espérance de retour.

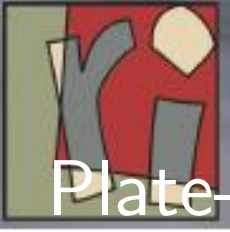
Approche



Défis

Malédiction de la dimensionalité

Horizon fini – Rationalité limitée



OpenDP : Programmation dynamique approchée

Plate-forme `http://opendpsourceforge.net`

Optimisation

- Méthodes de gradient (BFGS)
- Algorithmes d'évolution (OpenBeagle & CMA-ES)
- Quasi-random methodes
- Derivation-free algs.

Apprentissage

- Weka
- Torch (SVM, NN. K-nn)
- Fast Artificial NN

Echantillonnage

- Méthodes à basse discrépance
- Méthodes à basse dispersion
- Echantillonnage actif

Le jeu de Go

Pourquoi les jeux ?

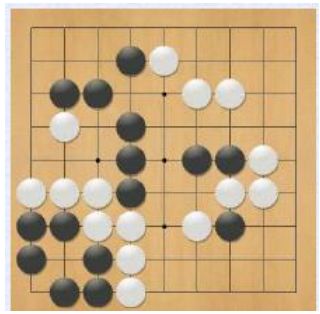
- Un entrainement à la vie (pour tous les mammifères...)
- Une simplification (les règles sont connues, le but aussi)
- Un test pour l'intelligence (artificielle)
- (Les jeux électroniques = 10 Milliards de $ aux USA en 2005)

Ce qui a été fait

- Jeu de Dames — Samuel, 1955
- Backgammon — Tesauro, 1992-1995
- Echecs — Deep Blue, 1997

Ce qui reste à faire

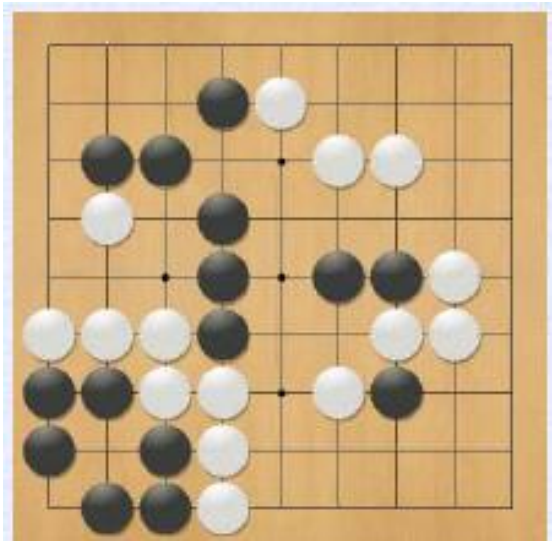
- Jeu de Go
- Jeu de Poker



Le défi du Go

Difficultés

- Nombre de positions légales 2.10^{170} (grand goban, 19x19)
  le nombre d'atomes dans l'univers...
- Nombre de coups possibles à chaque pas : en moyenne 200 (30-40 pour les échecs)
- Informations locales et globales (vie d'une chaine, symétries)
- Pas de fonction d'évaluation d'une position.



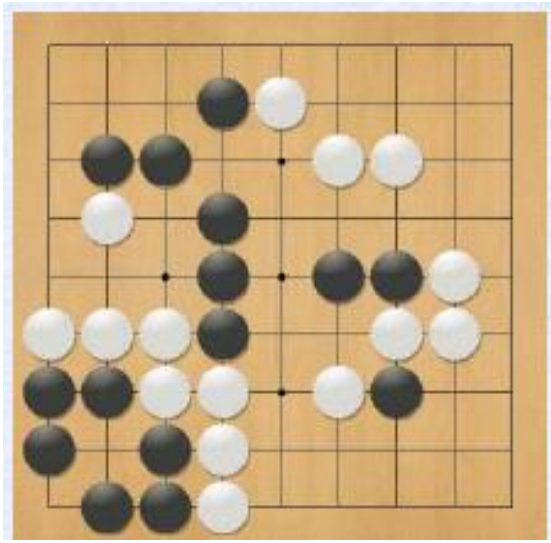
Les principes de MoGo

- Une évaluation aveugle *mais pas partiale*: le hasard
- La machine construit sa stratégie en son for intérieur

Une évaluation aveugle mais pas partiale

Evaluation Monte-Carlo Brügman (1993)

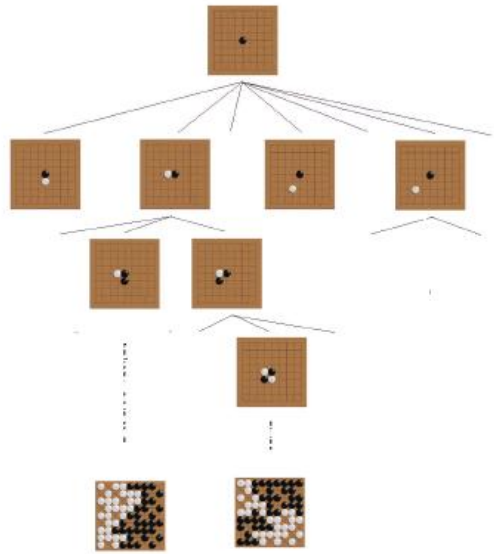
1. Ajouter aléatoirement des pierres blanches et noires
2. Jusqu'à remplir le goban
3. Mesurer qui gagne
4. Répéter 1-3 et prendre la moyenne



Remarque

- Cette heuristique est *faible*
- Mais uniformément faible.
- D'autres stratégies sont meilleures, avec des faiblesses ; or les faiblesses conduisent à attirer le jeu vers des régions où on croit qu'on gagne... et on perd

La machine construit sa stratégie : Monte-Carlo Tree Search



1. Considérer les coups possibles
2. Les évaluer en jouant contre un “partenaire fantome” (default partner)
3. Approfondir l’étude des meilleurs coups

La machine construit sa stratégie

Dilemme Explorer vs Exploiter Auer et al. 2001

Algorithme des Bandits Manchots

- Vous êtes dans un casino
- Vous voulez jouer aux machines à sous
- Vous voulez maximiser vos gains
- Vous devez jouer les meilleurs bras connus (exploitation)
- Vous devez explorer : peut-être y a-t-il un bras encore meilleur parmi ceux qu'on n'a pas essayés...

Algorithme des bandits manchots, 2

- Chaque coup possible est un bras
- Le gain associé est mesuré (Monte-Carlo)
- Il y a un intervalle de confiance (plus on l'a joué, moins il y a d'incertitudes sur le gain associé)
- Le principe est :

  Optimisme en face de l'incertain !

- Jouer le bras pour lequel

  $$\text{Gain} + \text{Incertitude}$$

  est maximum.

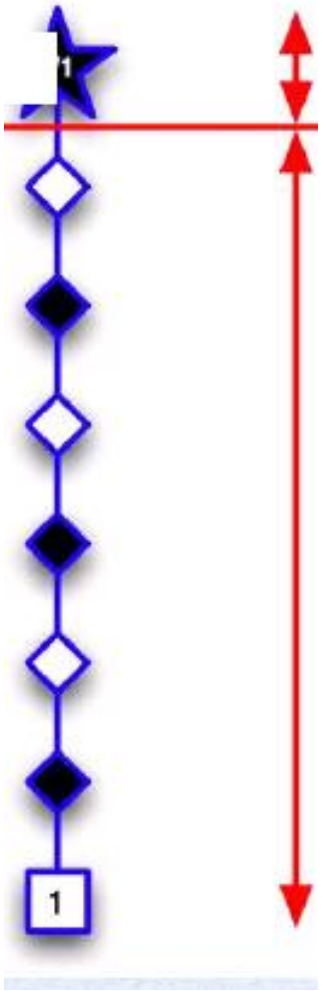
  maximiser Gain = favoriser l'exploitation

  maximiser Incertitude = favoriser l'exploration

$$\textit{Select argmax } \hat{\mu}_i + \sqrt{\frac{log(\sum n_j)}{n_i}}$$

Monte-Carlo Tree Search

Coup courant

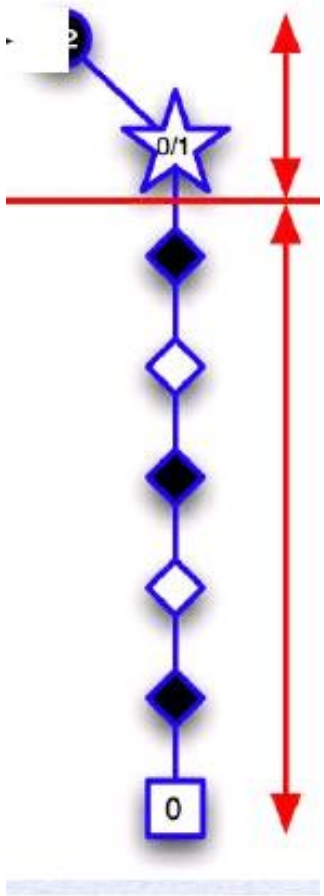


Stratégie en cours d'élaboration

Default partner

Monte-Carlo Tree Search

Coup courant

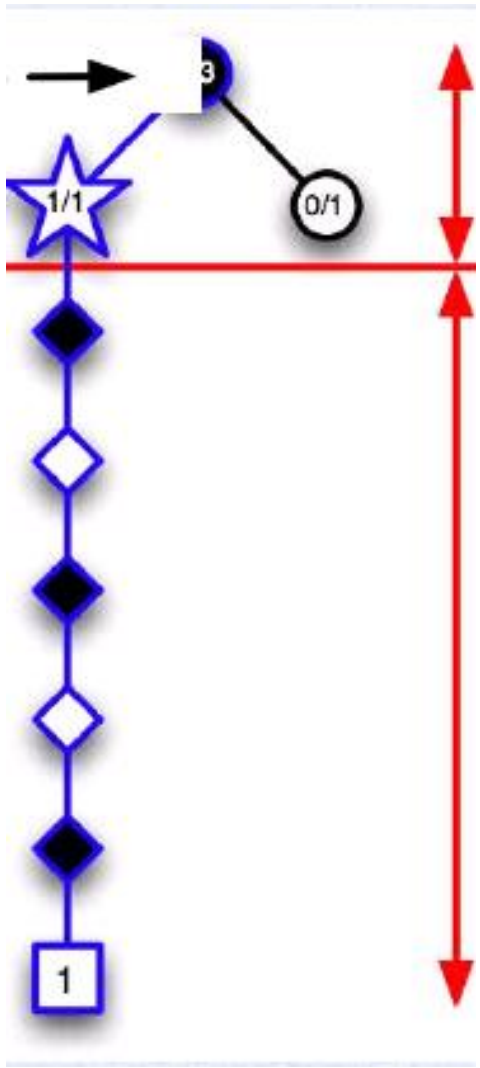


Stratégie en cours d'élaboration

Default partner

Monte-Carlo Tree Search

Coup courant

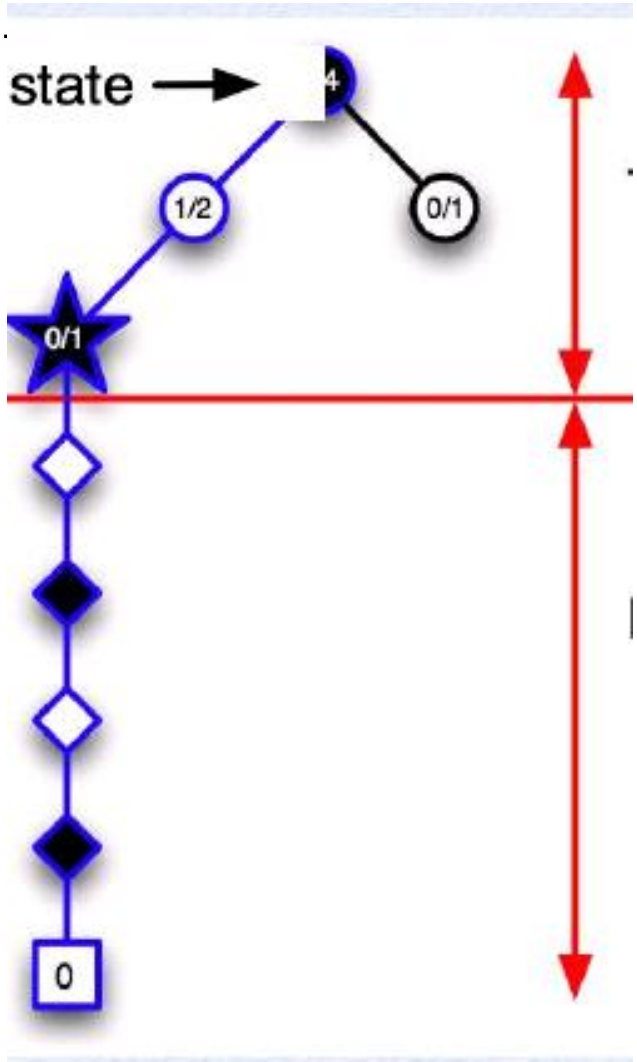


Stratégie en cours d'élaboration

Default partner

Monte-Carlo Tree Search

Coup courant



Stratégie en cours d'élaboration

Default partner

Autres aspects

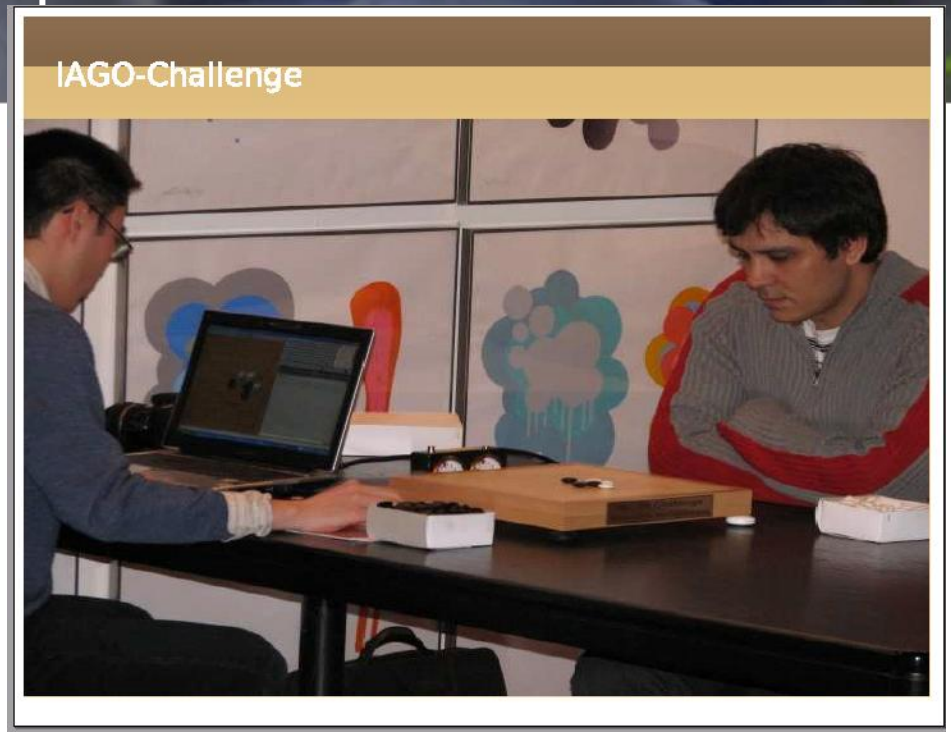
Force brute et parallélisation

- MoGo utilise un default partner qui DOIT etre rapide
- Il est préférable ici d'etre rapide et simple à plus intelligent et plus lent
- $\Rightarrow$ Parallélisation : BULL ; Grid 5000 ; Microsoft

Connaissances et sens commun (ici, prudence)

- Jouer près d'une pierre déjà jouée
- Identifier des motifs prioritaires (sauver une pierre qui va être prise)

Historique



- Médaille d'argent, Olypiades Beijing 2008
- Médaille d'or Taiwan (2008); Médaille d'or Hakone, Japon (2008)
- Première victoire en 19x19 handicap 9 contre un joueur 8e Dan humain;
- Seul programme à avoir gagné en 9x9 contre un joueur humain professionnel (mars 2008).

En guise de conclusion

Résumé

- <<A notable success of AI>> The Economist, Jan. 2007
- issu de la recherche fondamentale
- et d'une collaboration multi-tutelles et industrielle

Leçons

- Une alternative à la transmission d'expertise :
- La simulation pour l'acquisition d'expertise
- $\Rightarrow$ Des stratégies non humaines

Comment allier les deux mondes ?

Sylvain Gelly, Yizao Wang, Olivier Teytaud, Rémi Munos, Rémi Coulomb, Arpad Rimmel, Jean-Baptiste Hoock, Julien Perez, Thomas Hérault, Vincent Néri, Jean-Francois Méhaut, Jean-Yves Audibert