Reinforcement Learning

Michèle Sebag ; TP : Diviyan Kalainathan
TAO, CNRS − INRIA − Université Paris-Sud

Dec. 11th, 2017

Where we are

MDP Main Building block

General settings

	Model-based	Model-free
Finite	Dynamic Programming	**Discrete RL**
Infinite	(optimal control)	Continuous RL

More about the Exploration *vs* Exploitation Dilemma

This course: **Multi-Armed Bandits ; Monte-Carlo Tree Search**

Overview

Action selection as a Multi-Armed Bandit problem

Lai, Robbins 85

In a casino, one wants to maximize one's gains *while playing*.

Lifelong learning

Exploration vs **Exploitation** Dilemma

- Play the best arm so far ? **Exploitation**
- But there might exist better arms... **Exploration**

Formalization

- K options a.k.a. arms
- Arms are independent
- The i-th arm yields a reward r drawn iid along distribution ν_i
  In the following, $\nu_i = \text{Bernoulli}(\mu_i)$
  (return 1 with proba μ_i, 0 otherwise).

Goals

- Find the best arm:
$$i^* = \arg\max_i \mathbb{E}[\nu_i]$$
- Find a policy $\pi : t \rightarrow i_t$, gets reward r_t s.t. the sum of rewards is maximal in expectation
$$\pi = \arg\max \mathbb{E}[r_0 + r_1 + \ldots$$

Applications

- Find the best cure/drug for a disease.
  $r = 1$ if patient is cured, 0 otherwise
- Find the best ad for a Web site/user
  $r = 1$ if user clicks on the ad, 0 otherwise
- Find the best action for a robot
  $r = 1$ if the robot grasps the banana, 0 otherwise
  (What is different here ?)

The multi-armed bandit (MAB) problem

Algorithmic setting

Unknown parameters: K unknown probability distributions on $[0, 1]$

Known parameters: the set of arms $1 \dots K$, the number of rounds T

For each round $t = 1, 2, \dots, T$

(1) the learner chooses $i_t \in 1 \dots K$ according to its own strategy.

(2) the learner incurs and observes the reward $r_t \sim \nu_{i_t}$ independently from the past given rewards.

T: time horizon

When T unknown, algorithm is *anytime*

The multi-armed bandit (MAB) problem

- K arms
- Each arm gives reward 1 with probability μ_i, 0 otherwise
- Let $\mu^* = argmax\{\mu_1, \ldots \mu_K\}$, with $\Delta_i = \mu^* - \mu_i$
- In each time t, one selects an arm i_t and gets a reward r_t

$$
\begin{array}{lcll}
n_{i,t} & = & \sum_{u=1}^{t} \mathbb{1}_{i^*_u = i} & \textit{number of times } i \textit{ has been selected} \\
\hat{\mu}_{i,t} & = & \frac{1}{n_{i,t}} \sum_{i^*_u = i} r_u & \textit{average reward of arm } i
\end{array}
$$

Goal: Maximize $\sum_{u=1}^{t} r_u$

$\Leftrightarrow$

$$\textbf{Minimize Regret } (t) = \sum_{u=1}^{t} (\mu^* - r_u) = t\mu^* - \sum_{i=1}^{K} n_{i,t}\, \hat{\mu}_{i,t} \approx \sum_{i=1}^{K} n_{i,t} \Delta_i$$

Objective

Goal: Maximize $\sum_{u=1}^{t} r_u$

$\Leftrightarrow$

$$\textbf{Minimize Regret } (t) = \sum_{u=1}^{t} (r \sim \nu^* - r_u)$$

Regret: extra-loss incurred w.r.t. the oracle (who knows i^*).

Why using the regret ?

"Kind of" normalization w.r.t. problem difficulty: the more difficult the problem, the lower the oracle's gain; what matters is how well one fares compared to the expert.
(Additive normalization).

Overview

Notations

- $n_{i,t}$: number of times i has been selected up to t
- $\hat{\mu}_{i,t}$ empirical reward of i-th arm as of t

$$\hat{\mu}_{i,t} = \frac{1}{n_{i,t}} \sum_{u=1}^{t} r_u . \mathbb{I}_{i_u = i}$$

  with $\mathbb{I}_e = 1$ iff e holds true
- $\mu_i = \mathbb{E}[\nu_i]$
- Δ_i: margin of i-th arm

$$\Delta_i = \mu^* - \mu_i$$

Scientific questions

- How does the regret increase with T (linear ? quadratic ? logarithmic ?)
- What are the factors of difficulty of the MAB problem ?

Greedy algorithm

- Draw once each arm
$$\hat{\mu}_i = r \sim \nu_i$$
- At time u, select arm i_t s.t.
$$i_t = argmax\{\hat{\mu}_{i,t-1}, i = 1 \dots K\}$$

Example

- 2 arms:
  - arm 1, $\mu_1 = .8$;
  - arm 2, $\mu_2 = .2$.
- Assume the first two drawings yield:
  - arm 1, $r_1 = 0$;
  - arm 2, $r_2 = 1$.
- What happens ?

The ϵ-greedy algorithm

At each time t,

- With probability $1 - \varepsilon$
  select the arm with best empirical reward

$$i_t = argmax\{\hat{\mu}_{1,t}, \ldots \hat{\mu}_{K,t}\}$$

- Otherwise, select i_t uniformly in $\{1 \ldots K\}$

What is the regret ?

The ϵ-greedy algorithm

At each time t,

- With probability $1 - \varepsilon$
  select the arm with best empirical reward

$$i_t = argmax\{\hat{\mu}_{1,t}, \ldots \hat{\mu}_{K,t}\}$$

- Otherwise, select i_t uniformly in $\{1 \ldots K\}$

What is the regret ?

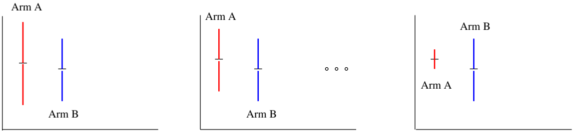
$$\textbf{Regret } (t) > \varepsilon t \frac{1}{K} \sum_i \Delta_i$$

But: Optimal regret rate: $log(t)$ Lai Robbins 85

Upper Confidence Bound

Auer et al. 2002

$$\text{Select } i_t = \text{argmax} \left\{ \hat{\mu}_{i,t} + \sqrt{2 \frac{log(t)}{n_{i,t}}} \right\}$$



Decision: Optimism in front of unknown !

Upper Confidence bound, 2

Thm: UCB achieves the optimal regret rate $log(t)$

$$\text{If } i_t = \text{ argmax } \left\{ \hat{\mu}_{i,t} + \sqrt{c_e \frac{log(\sum n_{j,t})}{n_{i,t}}} \right\}$$

Then

$$Regret(t) \leq 8 \sum_{i \neq i*} \frac{1}{\Delta_i} log(t) + \left(1 + \frac{\pi^2}{3}\right) \sum_i \Delta_i$$

Proof

$$Regret(t) = \sum_{i \neq i*} n_{i,t} \Delta_i$$

Upper Confidence bound, 3

The very useful Hoeffding inequality

Given $r_1, \dots r_n$ iid in $[0,1]$ drawn after p, with expectation μ,
Define empirical mean $\hat{\mu}_n = 1/n \sum_{u=1}^{n} r_u$, then

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu}_n - \mu \geq \varepsilon\right) &\leq \exp\left(-2\,\varepsilon^2 n\right), \\
\mathbb{P}\left(\mu - \hat{\mu}_n \geq \varepsilon\right) &\leq \exp\left(-2\,\varepsilon^2 n\right), \\
\mathbb{P}\left(|\hat{\mu}_n - \mu| \geq \varepsilon\right) &\leq 2\exp\left(-2\,\varepsilon^2 n\right)
\end{aligned}$$

Upper Confidence bound, 4

Sketch of the proof

Bound the number of times i is selected instead of i^*. This happens at step u iff

$$\hat{\mu}_{i,u} + \sqrt{\frac{2log(t)}{n_{i,u}}} > \hat{\mu}_{*,u} + \sqrt{\frac{2log(t)}{n_{*,u}}}$$

And we know that

$$\mu_* = \mu_i + \Delta_i$$

(a) Either $\hat{\mu}_{i,u}$ is close to μ_i

(b) Or $\hat{\mu}_{*,u}$ is close to μ_*

(c) Or, (a) and (b) are false, but this happens rarely (logarithmically...)

Upper Confidence bound, 5

$$\hat{\mu}_{i,u} + \sqrt{2\frac{log(t)}{n_{i,t}}} > \hat{\mu}_{*,u} + \sqrt{2\frac{log(t)}{n_{*,t}}}$$

One of the three equations below holds wrong

(a) $\hat{\mu}_{i,u} > \mu_i + \sqrt{2\frac{log(t)}{n_{i,t}}}$

(b) $\hat{\mu}_{*,u} < \mu_* - \sqrt{2\frac{log(t)}{n_{*,t}}}$

(c) $\sqrt{2\frac{log(t)}{n_{i,t}}} + \sqrt{2\frac{log(t)}{n_{*,t}}} > \Delta_i \Rightarrow$

$$n_{i,t}, n_{*,t} < \frac{8log(t)}{\Delta_i^2}$$

Decompose time: before and after step ℓ

With

$$\ell = \frac{8log(t)}{\Delta_i^2}$$

After ℓ, (c) is true; hence either (a) or (b) is wrong.

Upper Confidence bound, 6

$$n_{i,t} < \ell + \sum_{u=\ell}^{t} \mathbb{1}\{i_u = i\}$$

As $\ell = \frac{8 log(t)}{\Delta_i^2}$, either $\hat{\mu}_{i,u}$ or $\hat{\mu}_{*,u}$ is outside its confidence interval.
Hoeffding inequality yields (event (a)):

$$\Pr\left(\hat{\mu}_{i,t} - \mu_i \geq \sqrt{2\frac{log(t)}{n_{i,t}}}\right) \leq t^{-4}$$

Therefore (union bound)

$$\sum_{u=\ell}^{\infty} \mathbb{1}_{(a)} \leq \sum_{u=\ell}^{\infty} u^{-4}$$

Upper Confidence bound, 7

Known

$$\sum_{k=1}^{\infty} \frac{1}{k^{-4}} = (1 + \frac{\pi^2}{3})$$

Finally

$$\mathbb{E}[n_{i,t}\Delta_i] \leq \frac{8 log(t)}{\Delta_i^2} \times \Delta_i + (1 + \frac{\pi^2}{3})\Delta_i$$

QED: UCB regret is logarithmic

$$Regret(t) \leq 8 \sum_{i \neq i*} \frac{1}{\Delta_i} log(t) + \left(1 + \frac{\pi^2}{3}\right) \sum_i \Delta_i$$

Overview

Around MAB algorithms

- UCB is great, but not optimal. See KL-UCB Garivier et al. 2012
- In practice, play with C. control the exploration/exploitation trade-off
- Take into account the standard deviation of $\hat{\mu}_i$: Select $i_t = \text{argmax}$

$$\left\{ \hat{\mu}_{i,t} + \sqrt{c_e \frac{log(\sum n_{j,t})}{n_{i,t}} + min\left(\frac{1}{4}, \hat{\sigma}^2_{i,t} + \sqrt{c_e \frac{log(\sum n_{j,t})}{n_{i,t}}}\right)} \right\}$$

- When there are **many** arms: tendency to over-explore...

Extensions

- When there is some side information: contextual bandits
- When arm distributions are not stationary: restless bandits

A particular algorithm: BESA

Best Empirical Sampled Average Baransi Maillard 2014

Intuition

- Case 1: you compare two arms with same number of reward samples.
  Easy: take the one with best average.

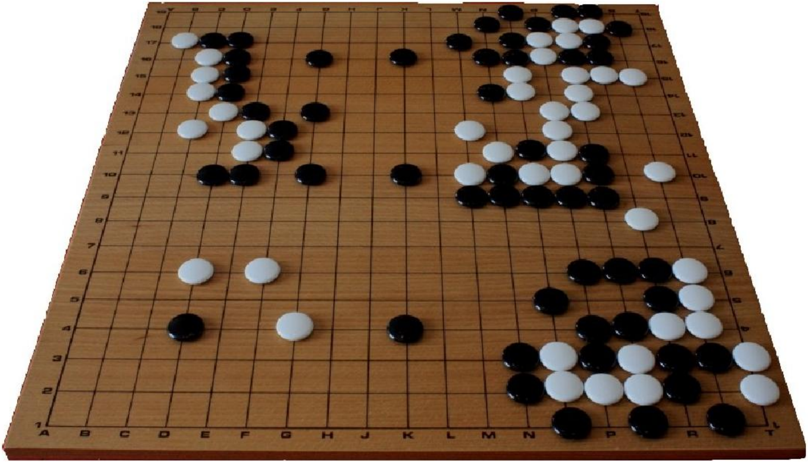
- Case 2: there is an arm A with many samples, and an arm B with few samples (say k).
  Easy: subsample k rewards for arm A and get back to Case 1.

Nota-bene

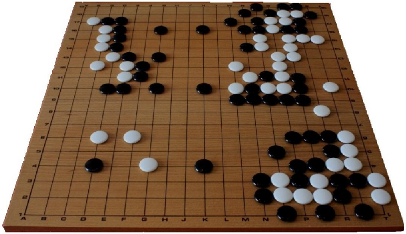
Same results with one hyper-parameter less == much better.

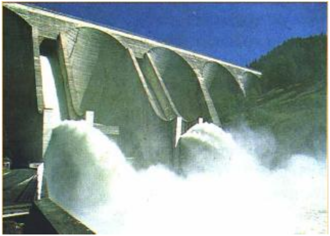
Overview

MCTS: computer-Go as explanatory example



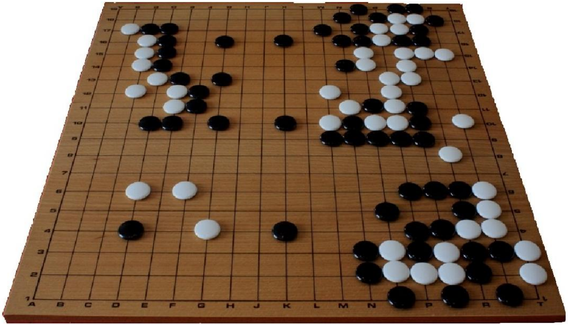
Not just a game: same approaches apply to optimal energy policy

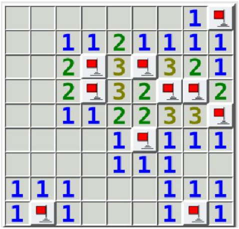




MCTS for computer-Go and MineSweeper

Go: deterministic transitions
MineSweeper: probabilistic transitions

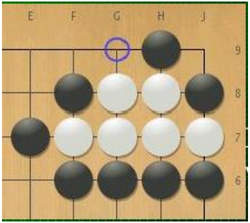




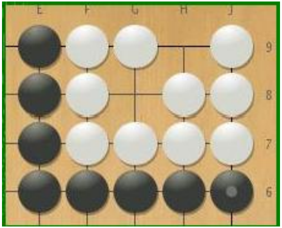
The game of Go in one slide

Rules

- Each player puts a stone on the goban, black first
- Each stone remains on the goban, except:



group w/o degree freedom is killed



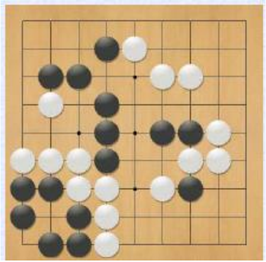
a group with two eyes can't be killed

- The goal is to control the max. territory

Go as a sequential decision problem

Features

- Size of the state space 2.10^{170}
- Size of the action space 200
- No good evaluation function
- Local and global features (symmetries, freedom, ...)
- A move might make a difference some dozen plies later



Setting

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Known transition model: $p(s, a, s')$
- Reward on final states: win or lose

Baseline strategies do not apply:

- Cannot grow the full tree
- Cannot safely cut branches
- Cannot be greedy

Monte-Carlo Tree Search

- An any-time algorithm
- Iteratively and asymmetrically growing a search tree

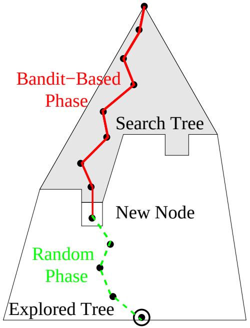
most promising subtrees are more explored and developed

Overview

Monte-Carlo Tree Search. Random phase

Gradually grow the search tree:

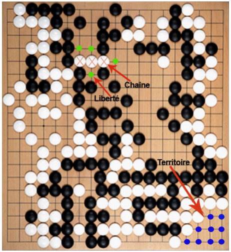
- Iterate Tree-Walk
  - Building Blocks
    - Select next action
      Bandit phase
    - Add a node
      Grow a leaf of the search tree
    - **Select next action bis**
      Random phase, roll-out
    - Compute instant reward
      Evaluate
    - Update information in visited nodes
      Propagate
- Returned solution:
  - Path visited most often



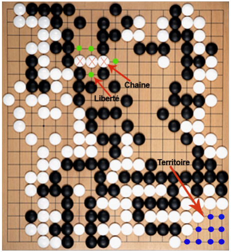
Random phase – Roll-out policy

Monte-Carlo-based Brügman 93

1. Until the goban is filled,
   add a stone (black or white in turn)
   at a uniformly selected empty position
2. Compute $r = \text{Win(black)}$
3. The outcome of the tree-walk is r



Random phase — Roll-out policy



Monte-Carlo-based Brügman 93

1. Until the goban is filled,
   add a stone (black or white in turn)
   at a uniformly selected empty position
2. Compute $r = \text{Win(black)}$
3. The outcome of the tree-walk is r

Improvements ?

- Put stones randomly in the neighborhood of a previous stone
- Put stones matching patterns — prior knowledge
- Put stones optimizing a value function — Silver et al. 07

Evaluation and Propagation

The tree-walk returns an evaluation r win(black)

Propagate

- For each node (s, a) in the tree-walk

$$\begin{aligned} n_{s,a} &\leftarrow n_{s,a} + 1 \\ \hat{\mu}_{s,a} &\leftarrow \hat{\mu}_{s,a} + \frac{1}{n_{s,a}}(r - \mu_{s,a}) \end{aligned}$$

Evaluation and Propagation

The tree-walk returns an evaluation r win(black)

Propagate

- For each node (s, a) in the tree-walk

$$\begin{array}{lcl} n_{s,a} & \leftarrow & n_{s,a} + 1 \\ \hat{\mu}_{s,a} & \leftarrow & \hat{\mu}_{s,a} + \frac{1}{n_{s,a}}(r - \mu_{s,a}) \end{array}$$

Variants

Kocsis & Szepesvári, 06

$$\hat{\mu}_{s,a} \leftarrow \begin{cases} min\{\hat{\mu}_x, x \text{ child of } (s,a)\} & \text{if } (s,a) \text{ is a black node} \\ max\{\hat{\mu}_x, x \text{ child of } (s,a)\} & \text{if } (s,a) \text{ is a white node} \end{cases}$$

Dilemma

- smarter roll-out policy $\rightarrow$
  more computationally expensive $\rightarrow$
  less tree-walks on a budget
- frugal roll-out $\rightarrow$
  more tree-walks $\rightarrow$
  more confident evaluations

Overview

Action selection revisited

$$\text{Select } a^* = \text{argmax} \left\{ \hat{\mu}_{s,a} + \sqrt{c_e \frac{log(n_s)}{n_{s,a}}} \right\}$$

- Asymptotically optimal
- But visits the tree infinitely often !

Being greedy is excluded not consistent

Frugal and consistent

$$\text{Select } a^* = \text{argmax} \frac{\text{Nb win}(s,a) + 1}{\text{Nb loss}(s,a) + 2}$$

Berthier et al. 2010

Further directions

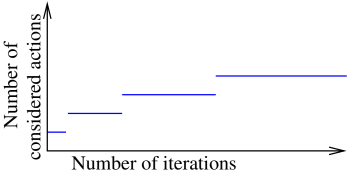
- Optimizing the action selection rule Maes et al., 11

Controlling the branching factor

What if many arms ? degenerates into exploration

- Continuous heuristics
  Use a small exploration constant c_e
- Discrete heuristics Progressive Widening
  Coulom 06; Rolet et al. 09

Limit the number of considered actions to $\lfloor \sqrt[b]{n(s)} \rfloor$
(usually $b = 2$ or 4)



Introduce a new action when $\lfloor \sqrt[b]{n(s)+1} \rfloor > \lfloor \sqrt[b]{n(s)} \rfloor$
(which one ? See RAVE, below).

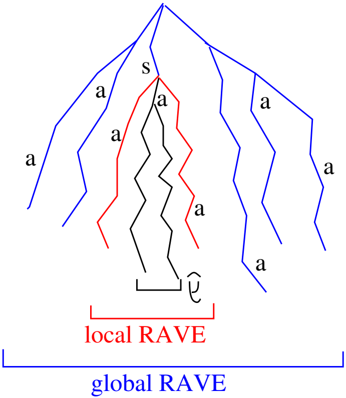
RAVE: Rapid Action Value Estimate

Gelly Silver 07

Motivation

- It needs some time to decrease the variance of $\hat{\mu}_{s,a}$
- Generalizing across the tree ?

$RAVE(s,a) =$
average $\{\hat{\mu}(s',a), s \text{ parent of } s'\}$



Rapid Action Value Estimate, 2

Using RAVE for action selection

In the action selection rule, replace $\hat{\mu}_{s,a}$ by

$$\alpha\hat{\mu}_{s,a} + (1-\alpha)\left(\beta RAVE_{\ell}(s,a) + (1-\beta)RAVE_{g}(s,a)\right)$$

$\alpha = \frac{n_{s,a}}{n_{s,a}+c_1}$ $\beta = \frac{n_{parent(s)}}{n_{parent(s)}+c_2}$

Using RAVE with Progressive Widening

- PW: introduce a new action if $\lfloor \sqrt[b]{n(s)+1} \rfloor > \lfloor \sqrt[b]{n(s)} \rfloor$
- Select promising actions: it takes time to recover from bad ones
- Select argmax $RAVE_{\ell}(parent(s))$.

A limit of RAVE

- Brings information from bottom to top of tree
- Sometimes harmful:

B2 is the only good move for white
B2 only makes sense as first move (not in subtrees)
⇒ RAVE rejects B2.

Improving the roll-out policy π

π_0 Put stones uniformly in empty positions

π_{random} Put stones uniformly in the neighborhood of a previous stone

π_{MoGo} Put stones matching patterns — prior knowledge

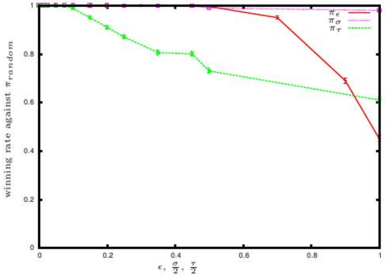
π_{RLGO} Put stones optimizing a value function — Silver et al. 07

Beware! — Gelly Silver 07

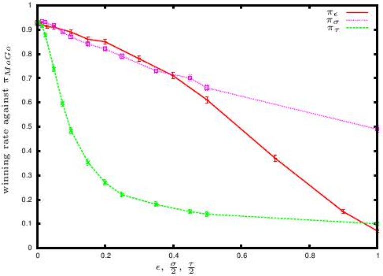
$$\pi \text{ better } \pi' \quad \not\Rightarrow \quad MCTS(\pi) \text{ better } MCTS(\pi')$$

Improving the roll-out policy π, followed

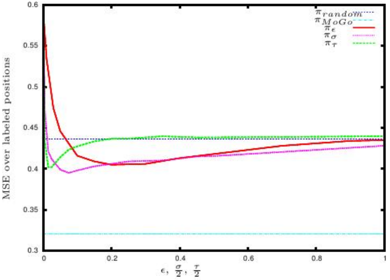
π_{RLGO} **against** π_{random}



π_{RLGO} **against** π_{MoGo}



Evaluation error on 200 test cases



Interpretation

What matters:

- Being **biased** is more harmful than being weak...
- Introducing a stronger but biased rollout policy π is detrimental.

if there exist situations where you (wrongly) think you are in good shape
then you go there
and you are in bad shape...

Using prior knowledge

Assume a value function $Q_{prior}(s,a)$

- Then when action a is first considered in state s, initialize

$$\begin{aligned} n_{s,a} &= n_{prior}(s,a) \quad \text{equivalent experience / confidence of priors} \\ \mu_{s,a} &= Q_{prior}(s,a) \end{aligned}$$

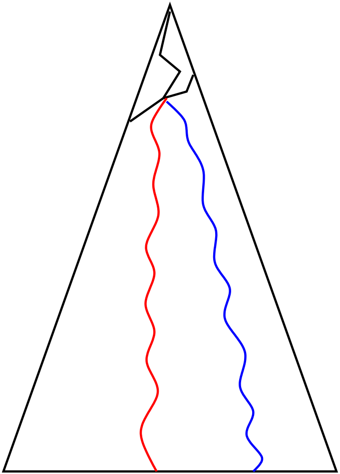
The best of both worlds

- Speed-up discovery of good moves
- Does not prevent from identifying their weaknesses

Overview

Parallelization. 1 Distributing the roll-outs

comp.
node 1

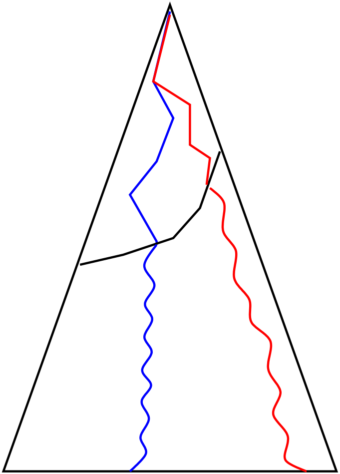


comp
node k

Distributing roll-outs on different computational nodes does not work.

Parallelization. 2 With shared memory

comp.
node 1

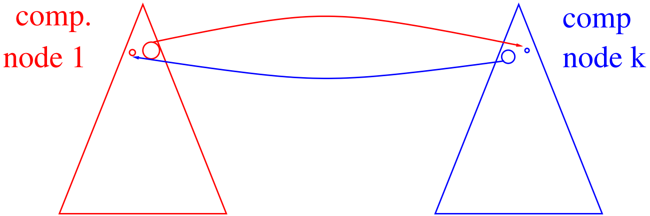


comp
node k

- Launch tree-walks in parallel on the same MCTS
- (micro) lock the indicators during each tree-walk update.

Use virtual updates to enforce the diversity of tree walks.

Parallelization. 3. Without shared memory



- Launch one MCTS per computational node
- k times per second $\quad k = 3$
  - Select nodes with sufficient number of simulations $> .05 \times \#$ total simulations
  - Aggregate indicators

Good news

Parallelization with and without shared memory can be combined.

It works !

32 cores against	Winning rate on 9×9	Winning rate on 19×19
1	75.8 ± 2.5	95.1 ± 1.4
2	66.3 ± 2.8	82.4 ± 2.7
4	62.6 ± 2.9	73.5 ± 3.4
8	59.6 ± 2.9	63.1 ± 4.2
16	$52\pm 3.$	63 ± 5.6
32	$48.9\pm 3.$	48 ± 10

Then:

- Try with a bigger machine ! and win against top professional players !
- Not so simple... there are diminishing returns.

Increasing the number N of tree-walks

N	$2N$ against N	
	Winning rate on 9×9	Winning rate on 19×19
1,000	71.1 ± 0.1	90.5 ± 0.3
4,000	68.7 ± 0.2	$84.5 \pm 0,3$
16,000	66.5 ± 0.9	80.2 ± 0.4
256,000	$61 \pm 0,2$	58.5 ± 1.7

The limits of parallelization

R. Coulom

Improvement in terms of performance against humans

$$\ll$$

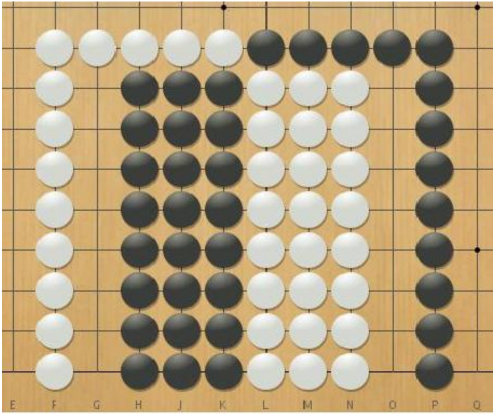
Improvement in terms of performance against computers

$$\ll$$

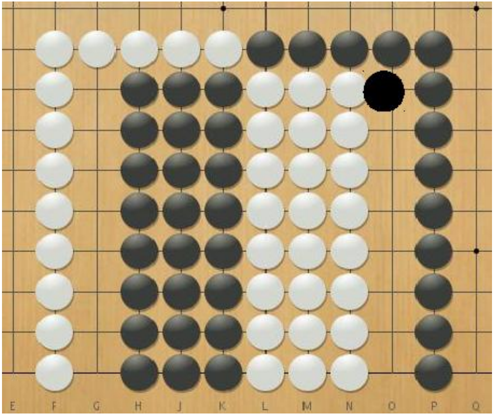
Improvements in terms of self-play

Overview

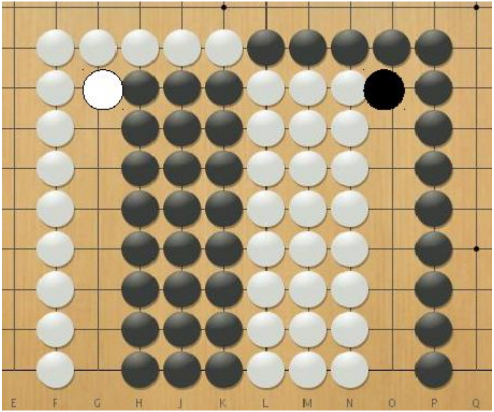
Failure: Semeai



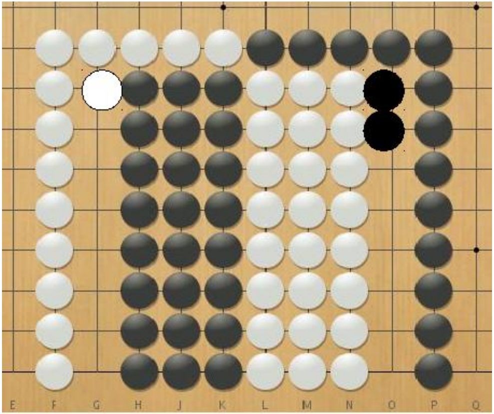
Failure: Semeai



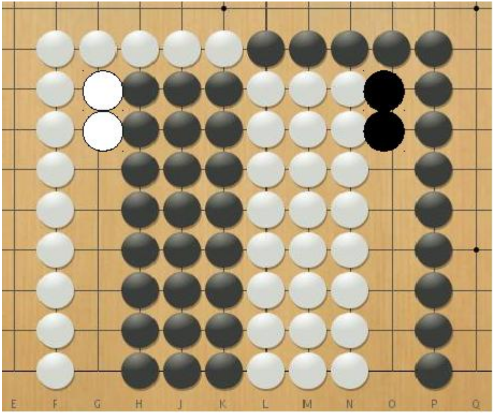
Failure: Semeai



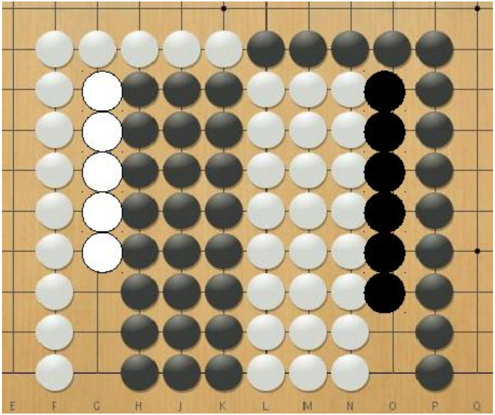
Failure: Semeai



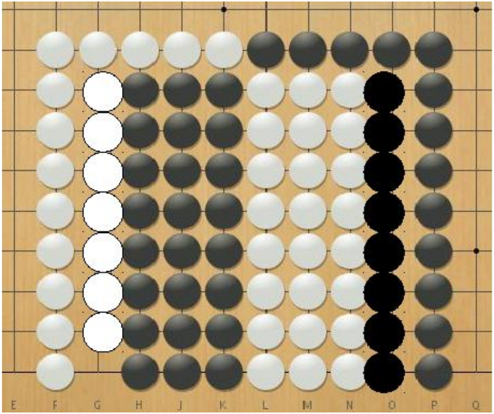
Failure: Semeai



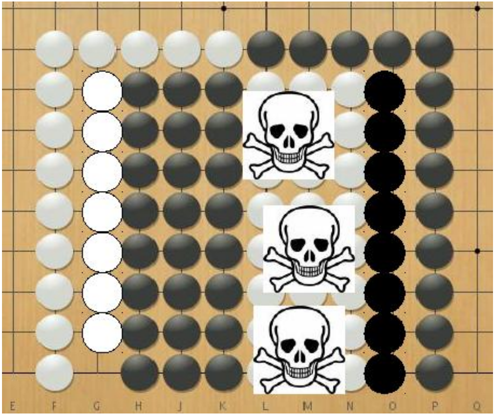
Failure: Semeai



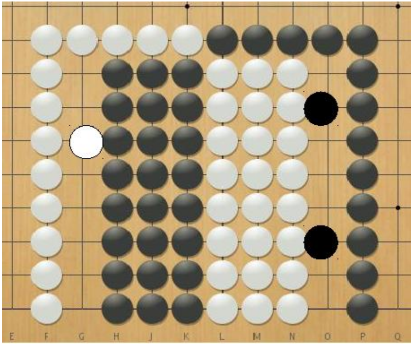
Failure: Semeai



Failure: Semeai



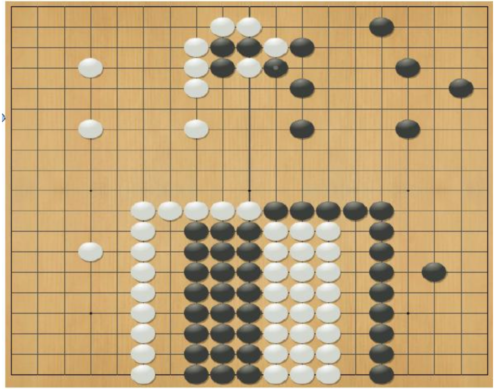
Failure: Semeai



Why does it fail

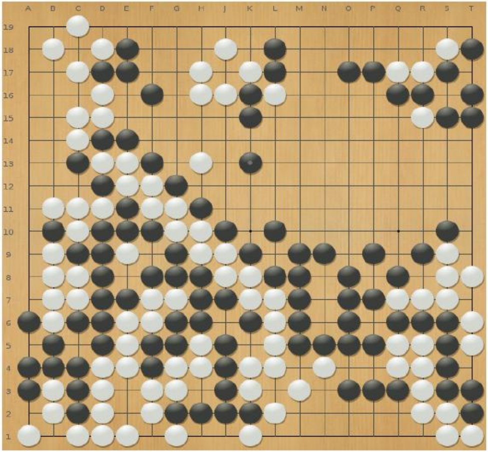
- First simulation gives 50%
- Following simulations give 100% or 0%
- But MCTS tries other moves: doesn't see all moves on the black side are equivalent.

Implication 1



MCTS does not detect invariance $\rightarrow$ too short-sighted and parallelization does not help.

Implication 2

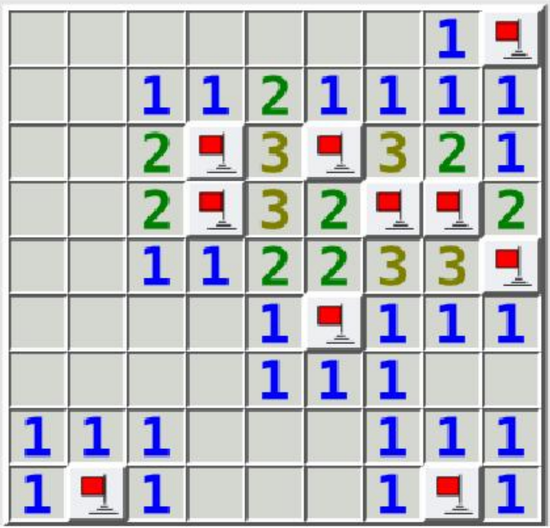


MCTS does not build abstractions $\rightarrow$ too short-sighted and parallelization does not help.

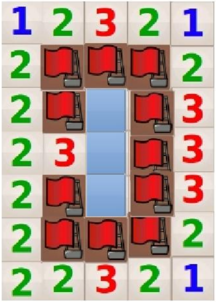
Overview

MCTS for one-player game

- The MineSweeper problem
- Combining CSP and MCTS

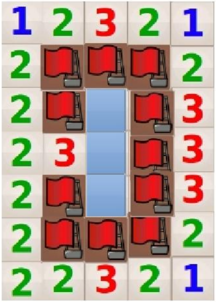


Motivation



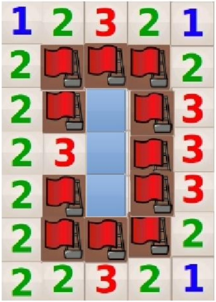
- All locations have same probability of death $1/3$
- Are then all moves equivalent ?

Motivation



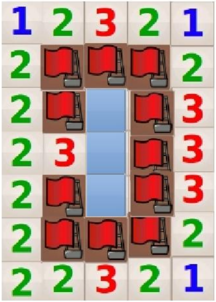
- All locations have same probability of death 1/3
- Are then all moves equivalent ? **NO !**

Motivation



- All locations have same probability of death 1/3
- Are then all moves equivalent ? **NO !**
- Top, Bottom: Win with probability 2/3

Motivation



- All locations have same probability of death $1/3$
- Are then all moves equivalent ? **NO !**
- Top, Bottom: Win with probability $2/3$
- MYOPIC approaches LOSE.

MineSweeper, State of the art

Markov Decision Process Very expensive; 4×4 is solved

Single Point Strategy (SPS) local solver

CSP

- Each unknown location j, a variable $x[j]$
- Each visible location, a constraint, e.g. $loc(15) = 4 \rightarrow$

$$x[04] + x[05] + x[06] + x[14] + x[16] + x[24] + x[25] + x[26] = 4$$

- Find all N solutions
- $\mathrm{P}(\text{mine in } j) = \frac{\text{number of solutions with mine in } j}{N}$
- Play j with minimal $\mathrm{P}(\text{mine in } j)$

Constraint Satisfaction for MineSweeper

State of the art

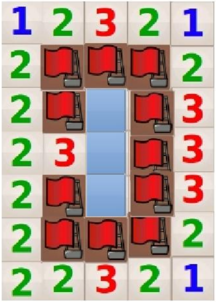
- 80% success *beginner* (9x9, 10 mines)
- 45% success *intermediate* (16x16, 40 mines)
- 34% success *expert* (30x40, 99 mines)

PROS

- Very fast

CONS

- Not optimal
- Beware of first move (opening book)



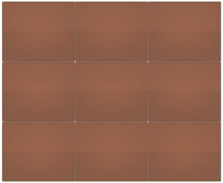
Upper Confidence Tree for MineSweeper

Couetoux Teytaud 11

- Cannot compete with CSP in terms of speed
- But consistent (find the optimal solution if given enough time)

Lesson learned

- Initial move matters
- UCT improves on CSP



- 3x3, 7 mines
- Optimal winning rate: 25%
- Optimal winning rate if uniform initial move: 17/72
- UCT improves on CSP by 1/72

UCT for MineSweeper

Another example

- 5x5, 15 mines
- GnoMine rule (first move gets 0)
- if 1st move is center, optimal winning rate is 100 %
- UCT finds it; CSP does not.

The best of both worlds

CSP

- Fast
- Suboptimal (myopic)

UCT

- Needs a generative model
- Asymptotic optimal

Hybrid

- UCT with generative model based on CSP

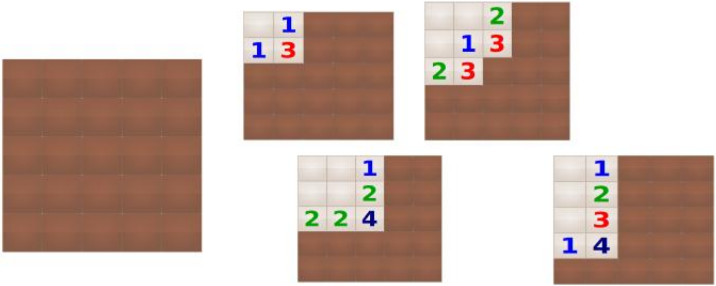
UCT needs a generative model

Given

- A state, an action
- **Simulate** possible transitions

Initial state, play top left

probabilistic transitions



Simulating transitions

- Using rejection (draw mines and check if consistent) SLOW
- Using CSP FAST

The algorithm: Belief State Sampler UCT

- One node created per simulation/tree-walk
- Progressive widening
- Evaluation by Monte-Carlo simulation
- Action selection: UCB tuned (with variance)
- Monte-Carlo moves
  - If possible, Single Point Strategy (can propose riskless moves if any)
  - Otherwise, move with null probability of mines (CSP-based)
  - Otherwise, with probability .7, move with minimal probability of mines (CSP-based)
  - Otherwise, draw a hidden state compatible with current observation (CSP-based) and play a safe move.

The results

- BSSUCT: Belief State Sampler UCT
- CSP-PGMS: CSP + initial moves in the corners

Format	CSP-PGMS	BSSUCT
4 mines on 4x4	64.7 %	**70.0% ± 0.6%**
1 mine on 1x3	100 %	100% (2000 games)
3 mines on 2x5	22.6%	**25.4 % ± 1.0%**
10 mines on 5x5	8.20%	9% (p-value: 0.14)
5 mines on 1x10	12.93%	**18.9% ± 0.2%**
10 mines on 3x7	4.50%	**5.96% ± 0.16%**
15 mines on 5x5	0.63%	**0.9% ± 0.1%**

Partial conclusion

Given a myopic solver

- It can be combined with MCTS / UCT:
- Significant (costly) improvements

Overview

Active Learning, position of the problem

Supervised learning, the setting

- Target hypothesis h^*
- Training set $\mathcal{E} = \{(x_i, y_i), i = 1 \dots n\}$
- Learn h_n from $\mathcal{E}$

Criteria

- Consistency: $h_n \rightarrow h^*$ when $n \rightarrow \infty$.
- Sample complexity: number of examples needed to reach the target with precision ϵ

$$\epsilon \rightarrow n_\epsilon \ s.t. \ ||h_n - h^*|| < \epsilon$$

Active Learning, definition

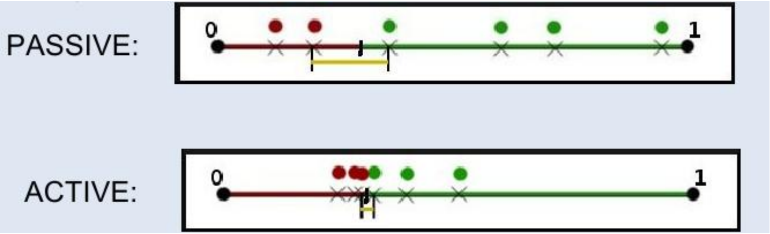
Passive learning iid examples

$$\mathcal{E} = \{(x_i, y_i), i = 1 \ldots n\}$$

Active learning

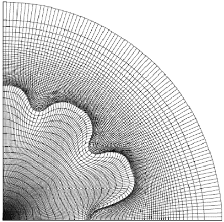
x_{n+1} selected depending on $\{(x_i, y_i), i = 1 \ldots n\}$

In the best case, exponential improvement:

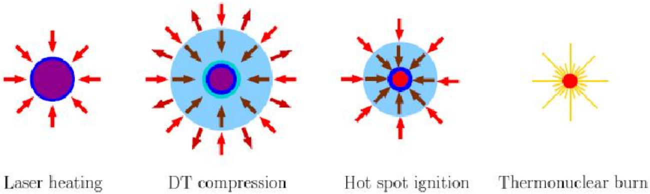


A motivating application

Numerical Engineering



- Large codes
- Computationally heavy $\sim$ days
- not fool-proof



Inertial Confinement Fusion, ICF

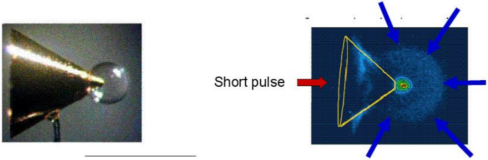
Goal

Simplified models

- Approximate answer
- … for a fraction of the computational cost
- Speed-up the design cycle
- Optimal design

More is Different

Alternative scheme : spherical target with a gold cone*



* Kodama et al. Nature **412** 798 (2001); **418** 933 (2002);

Active Learning as a Game

Ph. Rolet, 2010

Optimization problem

Find $F^* = argmin$
$\mathbb{E}_{h \sim \mathcal{A}(\mathcal{E},\sigma,T)} \mathbf{Err}(h, \sigma, T)$

$\mathcal{E}$: Training data set

$\mathcal{A}$: Machine Learning algorithm

$\mathcal{Z}$: Set of instances

$\sigma : \mathcal{E} \mapsto \mathcal{Z}$ sampling strategy

T: Time horizon

Err: Generalization error

Bottlenecks

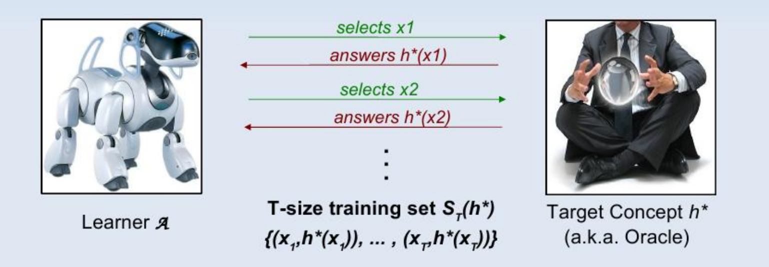
- Combinatorial optimization problem
- Generalization error unknown

Where is the game ?

- Wanted: a good strategy to find, as accurately as possible, the true target concept.
- If this is a game, you play it only once !
- But you can train...

Training game: Iterate

- Draw a possible goal (fake target concept h^*); use it as oracle
- Try a policy (sequence of instances $\mathcal{E}_{h^*,T} = \{(x_1, h^*(x_1)), \ldots (x_T, h^*(x_T))\}$
- Evaluate: Learn h from $\mathcal{E}_{h^*,T}$. Reward $= ||h - h^*||$



Overview

Conclusion

Take-home message: MCTS/UCT

- enables any-time smart look-ahead for better sequential decisions in front of uncertainty.
- is an integrated system involving two main ingredients:
  - Exploration vs Exploitation rule — UCB, UCBtuned, others
  - Roll-out policy
- can take advantage of prior knowledge

Caveat

- The UCB rule was not an essential ingredient of MoGo
- Refining the roll-out policy $\not\Rightarrow$ refining the system
  Many tree-walks might be better than smarter (biased) ones.

On-going, future, call to arms

Extensions

- Continuous bandits: action ranges in a $\mathbb{R}$ — Bubeck et al. 11
- Contextual bandits: state ranges in $\mathbb{R}^d$ — Langford et al. 11
- Multi-objective sequential optimization — Wang Sebag 12

Controlling the size of the search space

- Building abstractions
- Considering nested MCTS (partially observable settings, e.g. poker)
- Multi-scale reasoning

Bibliography

- Peter Auer, Nicolò Cesa-Bianchi, Paul Fischer: Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning 47(2-3): 235-256 (2002)
- Vincent Berthier, Hassen Doghmen, Olivier Teytaud: Consistency Modifications for Automatically Tuned Monte-Carlo Tree Search. LION 2010: 111-124
- Sébastien Bubeck, Rémi Munos, Gilles Stoltz, Csaba Szepesvári: X-Armed Bandits. Journal of Machine Learning Research 12: 1655-1695 (2011)
- Pierre-Arnaud Coquelin, Rémi Munos: Bandit Algorithms for Tree Search. UAI 2007: 67-74
- Rémi Coulom: Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search. Computers and Games 2006: 72-83
- Romaric Gaudel, Michèle Sebag: Feature Selection as a One-Player Game. ICML 2010: 359-366

- Sylvain Gelly, David Silver: Combining online and offline knowledge in UCT. ICML 2007: 273-280
- Levente Kocsis, Csaba Szepesvári: Bandit Based Monte-Carlo Planning. ECML 2006: 282-293
- Francis Maes, Louis Wehenkel, Damien Ernst: Automatic Discovery of Ranking Formulas for Playing with Multi-armed Bandits. EWRL 2011: 5-17
- Arpad Rimmel, Fabien Teytaud, Olivier Teytaud: Biasing Monte-Carlo Simulations through RAVE Values. Computers and Games 2010: 59-68
- David Silver, Richard S. Sutton, Martin Müller: Reinforcement Learning of Local Shape in the Game of Go. IJCAI 2007: 1053-1058
- Olivier Teytaud, Michèle Sebag: Combining Myopic Optimization and Tree Search: Application to MineSweeper, LION 2012.