Monte-Carlo Tree Search (MCTS) for Computer Go

Yang-Qiao Meng

University of Toronto

Nov 17th, 2016

Outline

- Background of Go
- MCTS
- UCT
- RAVE
- Heuristic MC-RAVE

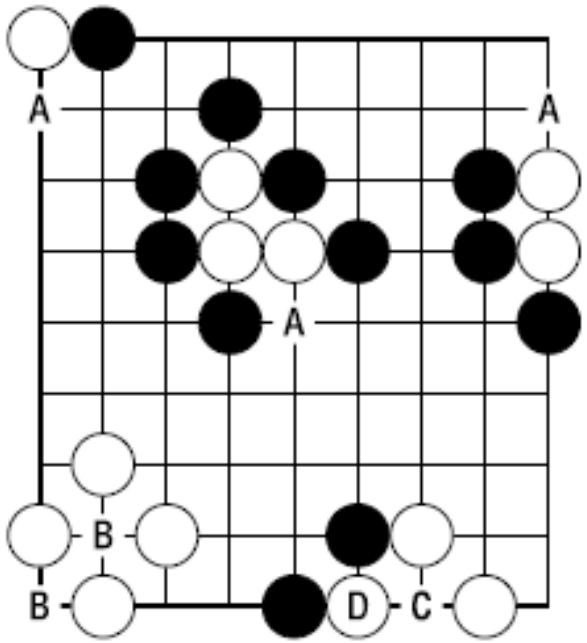
Acknowledgement

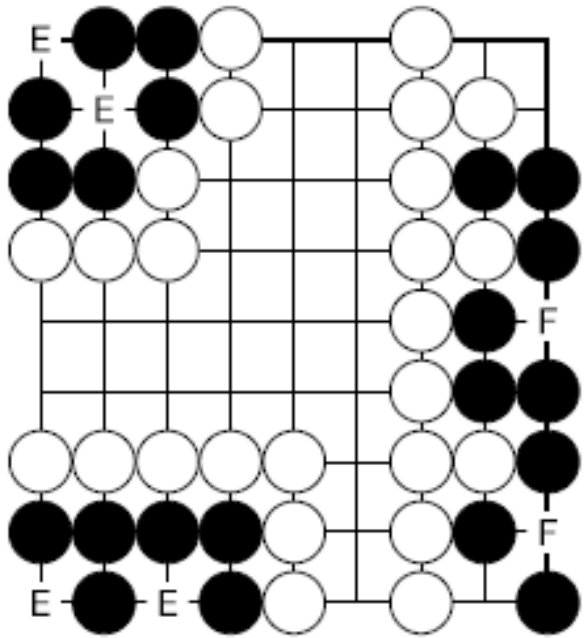
- Bruno Bouzy, Université Paris Descartes
- Adrien Couetoux, Martin Muller and Olivier Teytaud

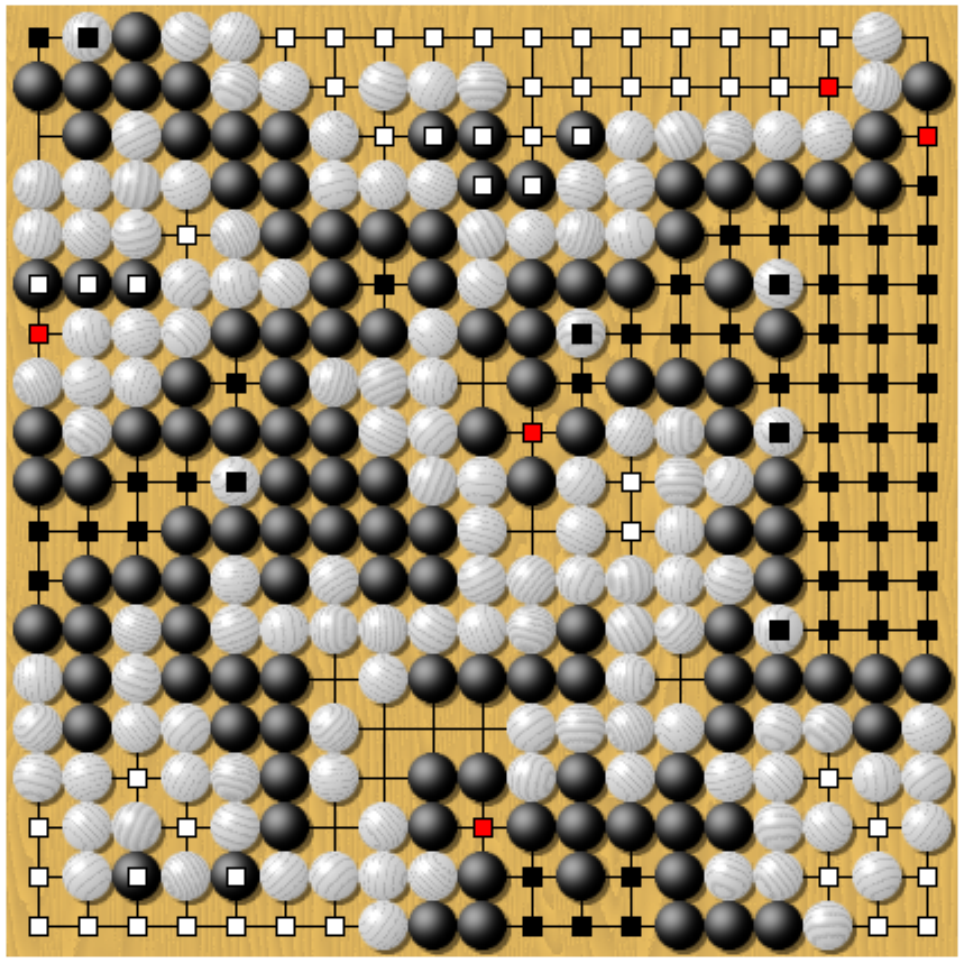
The game of Go

- Originate from China 4th century BC
- Korea and Japan, 5^{th}-7^{th} century CE
- 19 x 19 board
- 9 x 9 board for beginner
- State space: 10^170
- Rule: encirclement & occupation

Example







Monte Carlo Tree Search

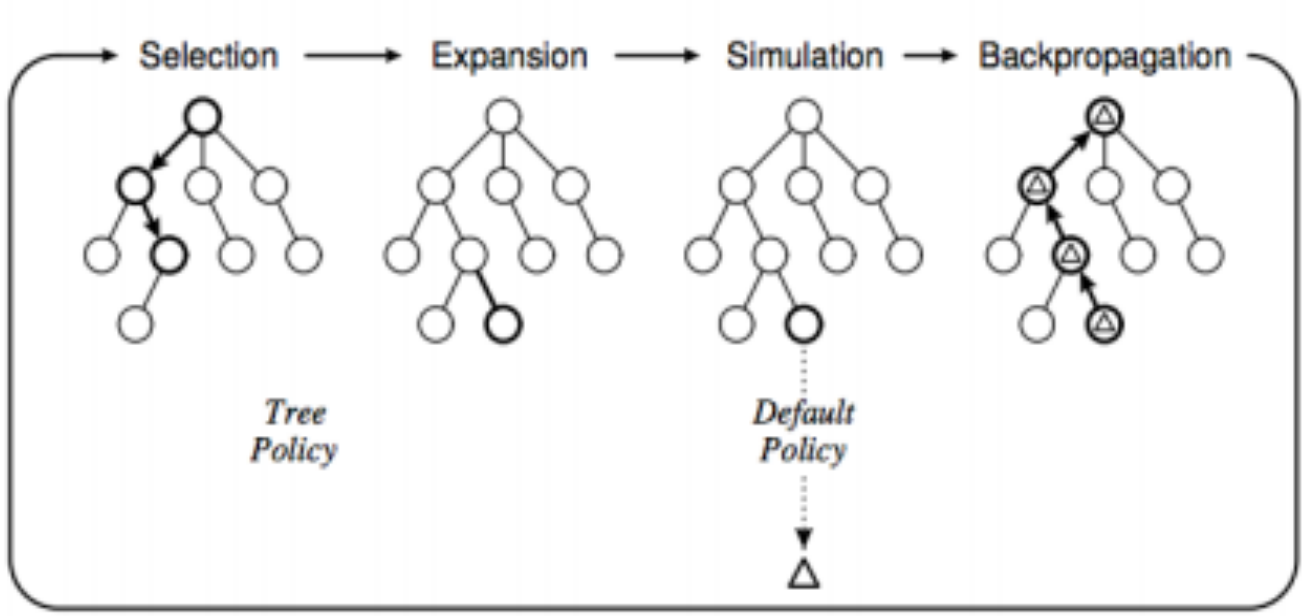
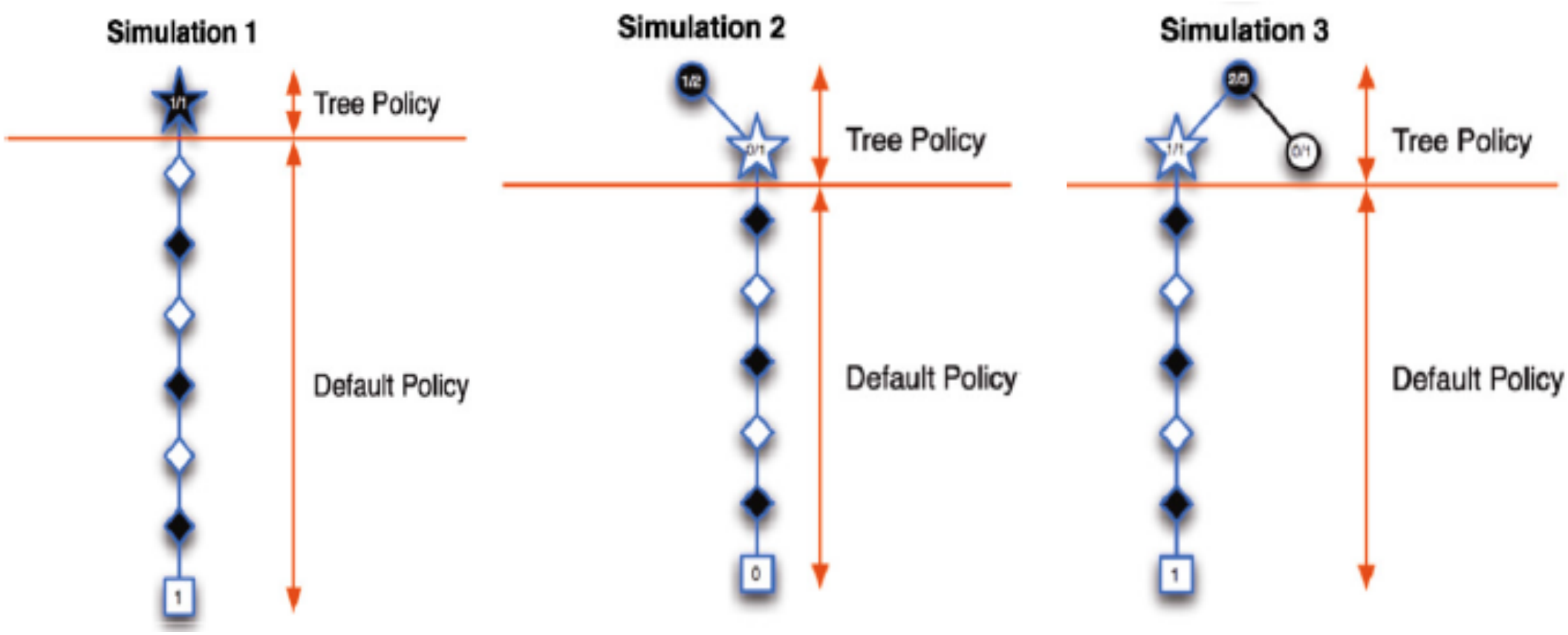
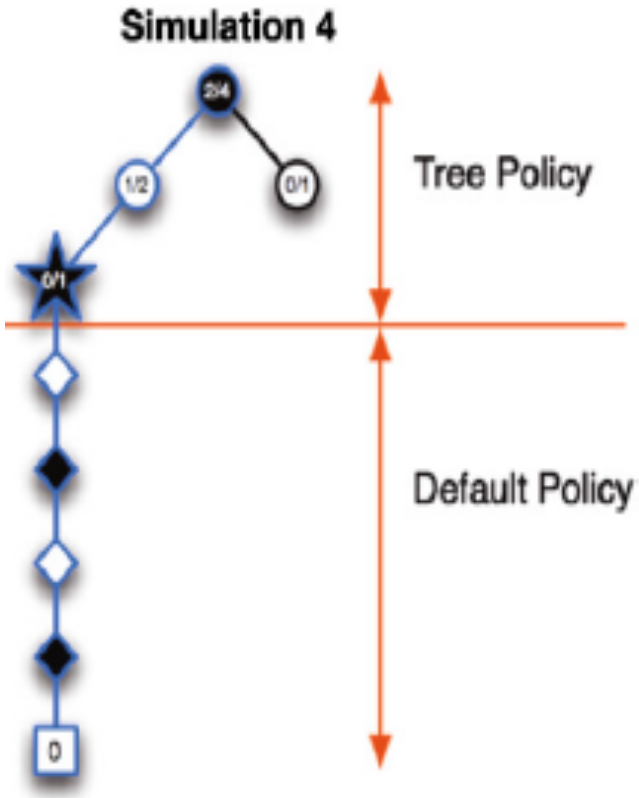


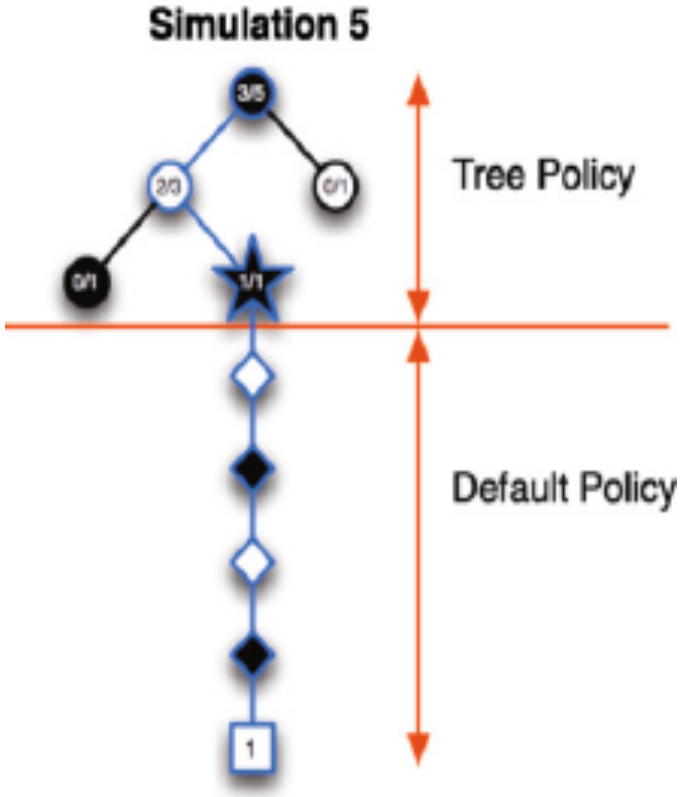
Fig. 4 The four steps of the main loop of MCTS. Each iteration of the loop corresponds to one possible sequence of states and moves, and adds information to the tree.

Monte Carlo Tree Search



Monte Carlo Tree Search





$$N(s_t) \leftarrow N(s_t) + 1,$$

$$N(s_t, a_t) \leftarrow N(s_t, a_t) + 1,$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \frac{z - Q(s_t, a_t)}{N(s_t, a_t)}$$

Upper Confidence Tree (UCT)

- Each state is a multi-arm bandit
- Each action is a bandit arm

$$Q^{\oplus}(s,a) = Q(s,a) + c\sqrt{\frac{\log N(s)}{N(s,a)}}$$

$$a^* = \underset{a}{\operatorname{argmax}}\, Q^{\oplus}(s,a)$$

Example

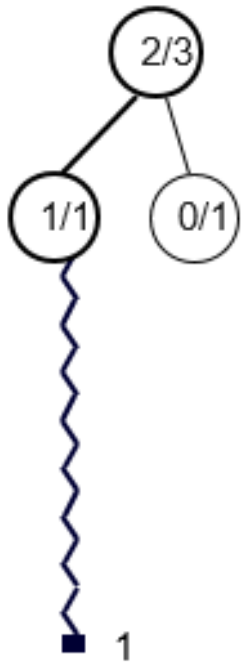
- 1 iteration

Example

- 2 iterations

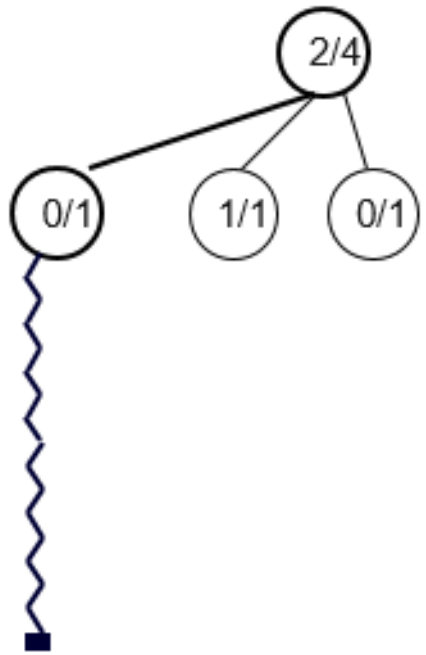
Example

- 3 iterations



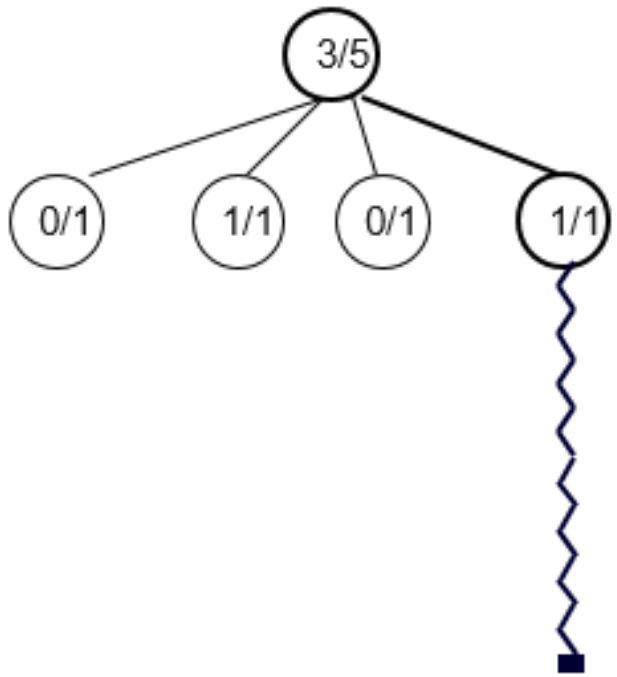
Example

- 4 iterations



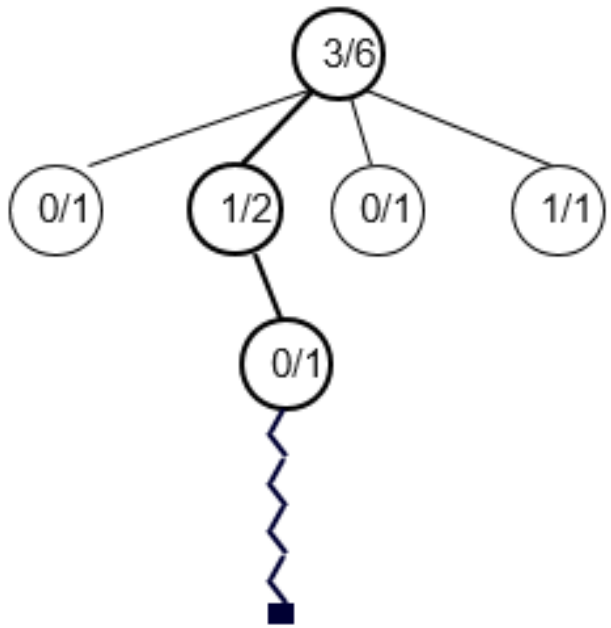
Example

- 5 iterations



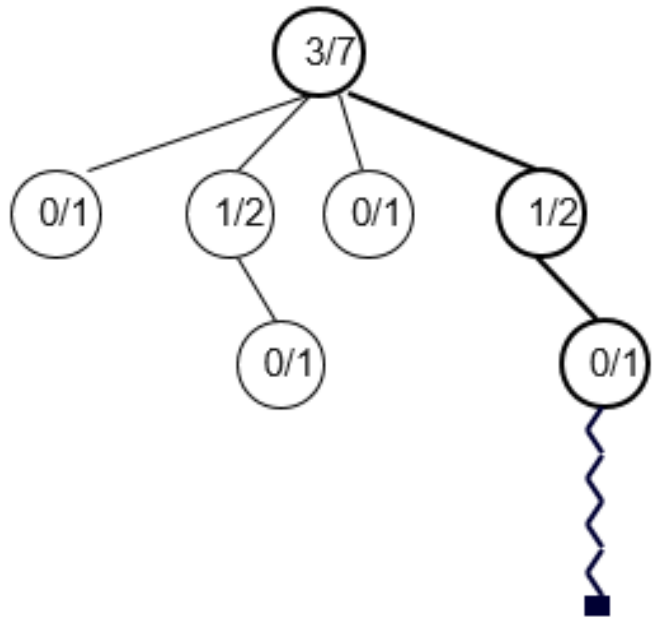
Example

- 6 iterations



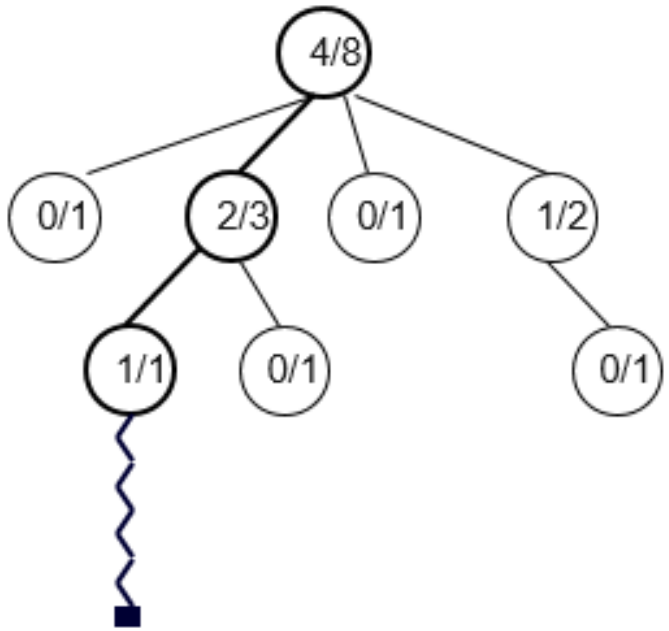
Example

- 7 iterations



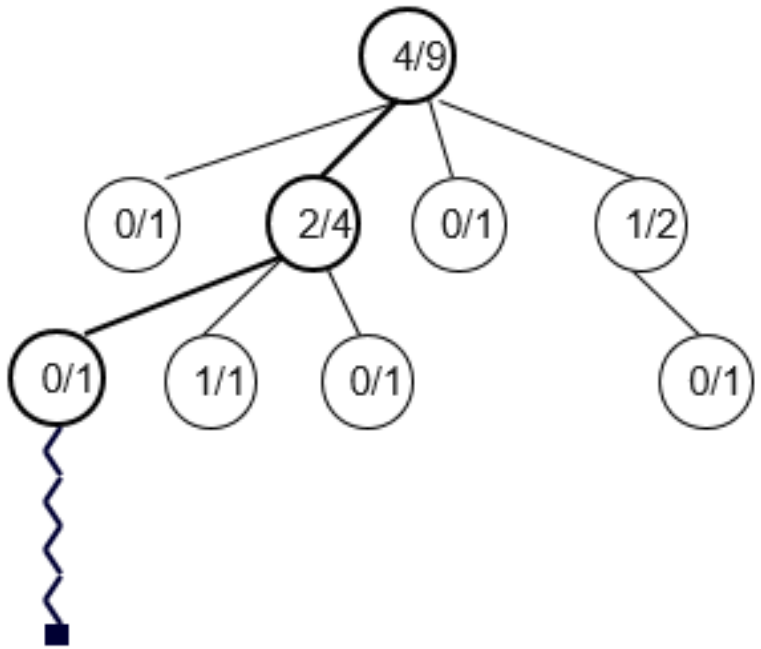
Example

- 8 iterations



Example

- 9 iterations



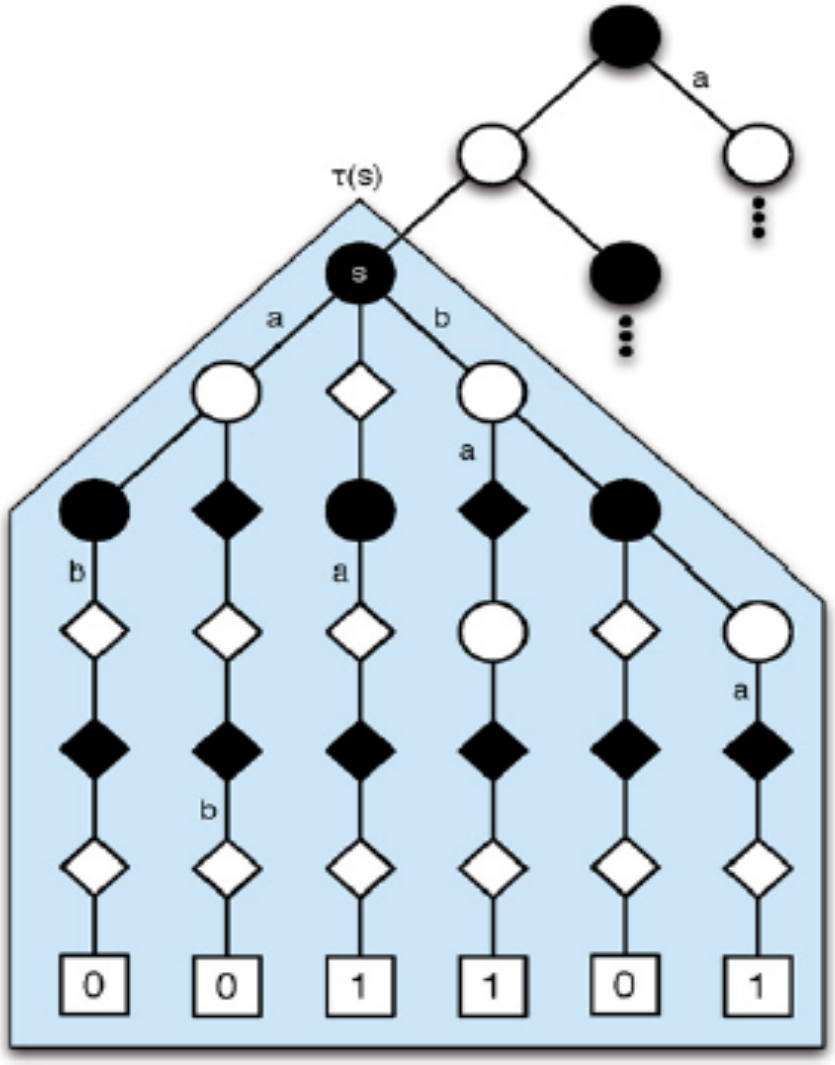
Rapid Action Value Estimation(RAVE)

- All-move-as-first(AMAF)
- A general value for each move, regardless when it is played

$$\tilde{Q}^{\pi}(s,a) = \mathbb{E}_{\pi}[z \mid s_t = s,\ \exists u \geqslant t \text{ s.t. } a_u = a].$$

$$\tilde{Q}^{\pi}(s,a) = Q^{\pi}(s,a) + \tilde{B}(s,a)$$

RAVE



$$Q(s,a) = 0/2$$
$$Q(s,b) = 2/3$$
$$\tilde{Q}(s,a) = 3/5$$
$$\tilde{Q}(s,b) = 2/5$$

RAVE

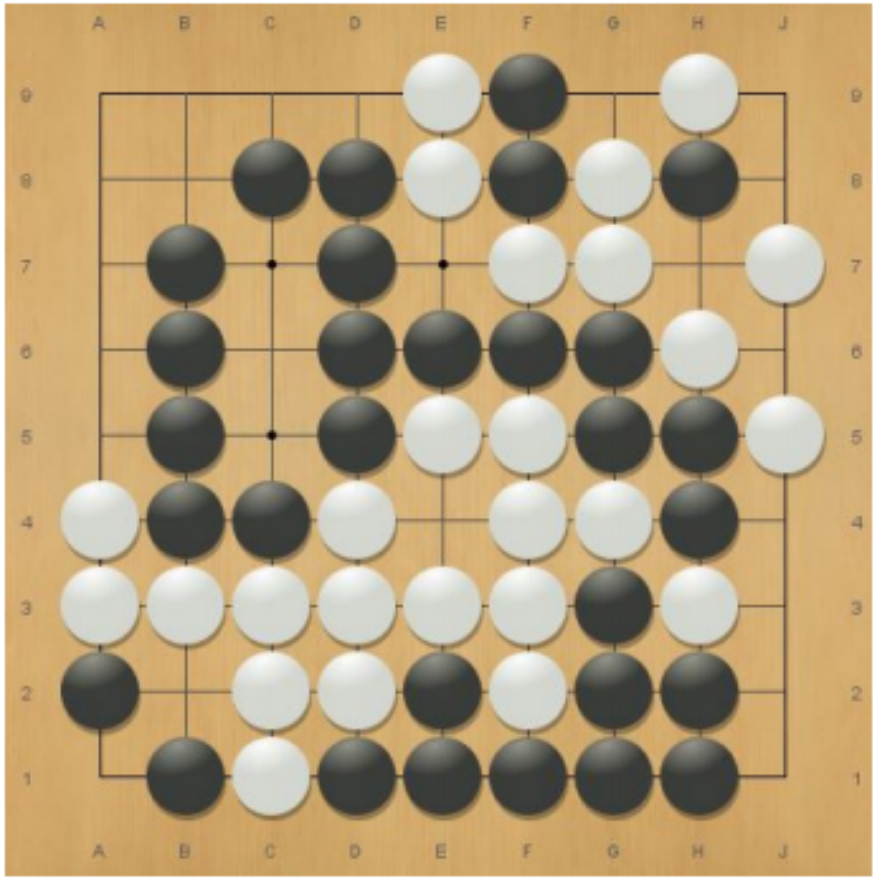


Fig. 5 Example of misleading RAVE score in FUEGO: White B2 is the only winning move which avoids a seki. Its RAVE value is very low, since if the liberty at A5 gets filled first, as happens in many simulations, then B2 becomes a very bad self-atari.

MC-RAVE

- Weighted sum between MC value and AMAF value

$$Q_{\star}(s,a) = \big(1 - \beta(s,a)\big) Q(s,a) + \beta(s,a) \tilde{Q}(s,a)$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \frac{z - Q(s_t, a_t)}{N(s_t, a_t)}$$

$$\tilde{Q}(s_t, a_u) \leftarrow \tilde{Q}(s_t, a_u) + \frac{z - \tilde{Q}(s_t, a_u)}{\tilde{N}(s_t, a_u)}.$$

UCT-RAVE

- MC + UCT + RAVE

$$Q_{\star}^{\oplus}(s,a) = Q_{\star}(s,a) + c\sqrt{\frac{\log N(s)}{N(s,a)}}.$$

Schedule

- Hand selected schedule

$$\beta(s,a) = \sqrt{\frac{k}{3N(s)+k}}$$

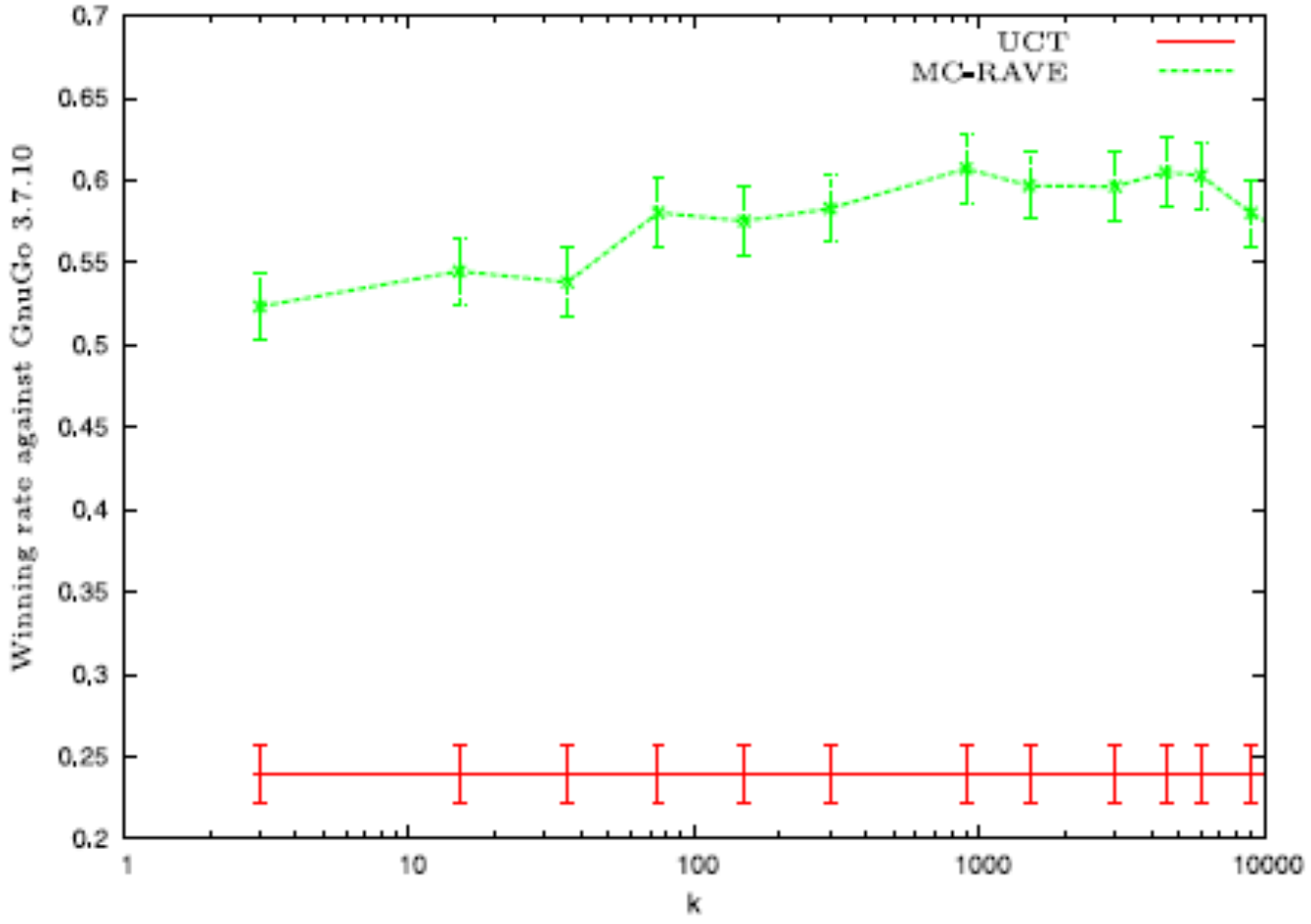
where k specifies the number of simulations at which the Monte-Carlo value and the AMAF value should be given equal weight, $\beta(s,a) = \frac{1}{2}$,

- Minimizing MSE schedule

$$e_\star^2 = \mathbb{E}\left[\left(Q_\star(s,a) - Q^\pi(s,a)\right)^2 \mid N(s,a) = n, \tilde{N}(s,a) = \tilde{n}\right]$$

$$\beta = \frac{\tilde{n}}{n + \tilde{n} + n\tilde{n}\tilde{b}^2/\mu_\star(1-\mu_\star)}$$

UCT vs MC-RAVE



Heuristic MC-RAVE

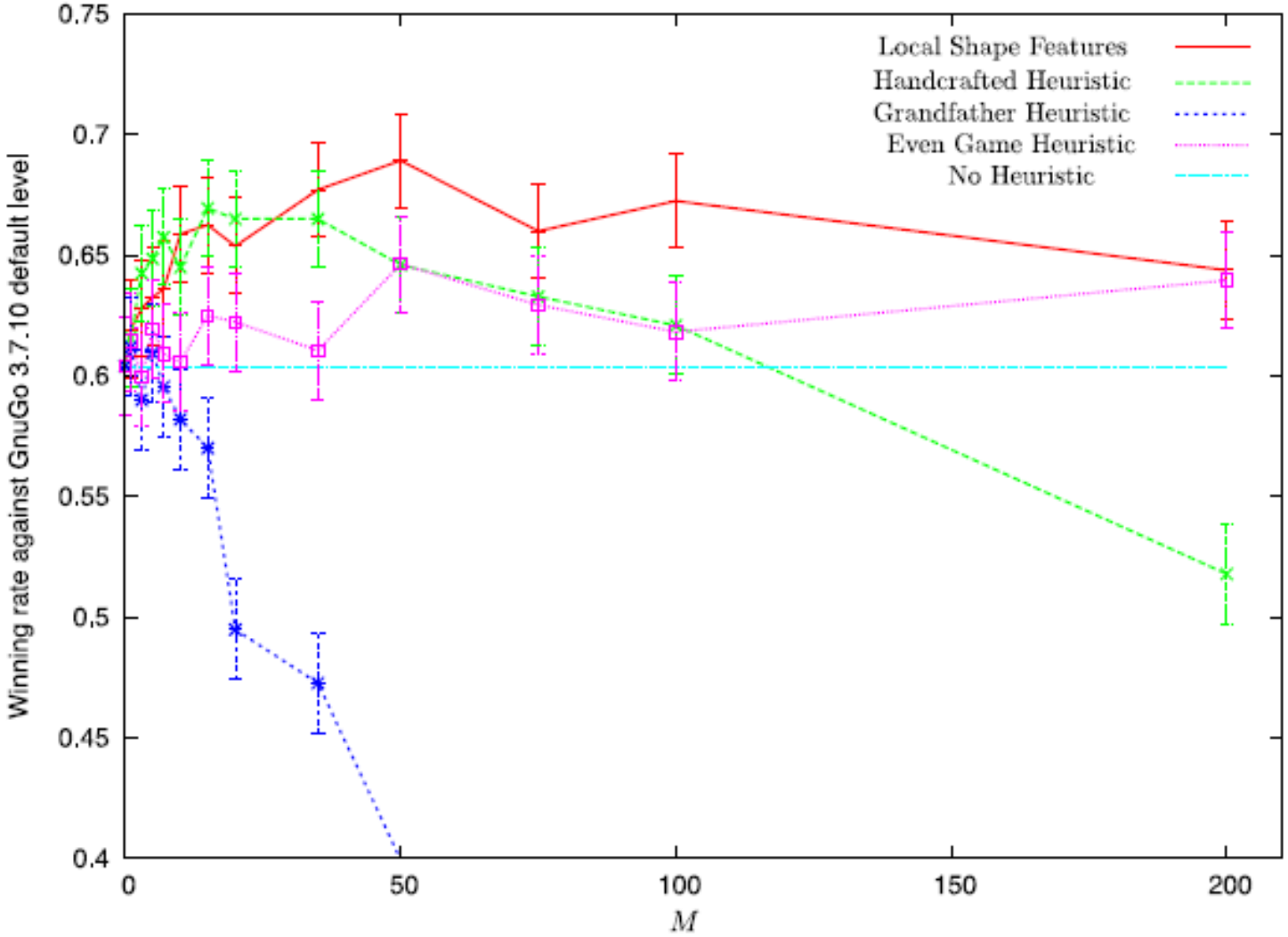
- Heuristic evaluation function: H(s, a)
- Heuristic confidence function: C(s ,a)

$$Q(s,a) \leftarrow H(s,a). \qquad \tilde{Q}(s,a) \leftarrow H(s,a)$$

$$N(s,a) \leftarrow C(s,a) \qquad \tilde{N}(s,a) \leftarrow \tilde{C}(s,a),$$

$$N(s) \leftarrow \sum_{a \in \mathcal{A}} N(s,a)$$

Heuristic MC-RAVE



Questions ?