Immediate Versus Delayed Rewards for the Game of Go

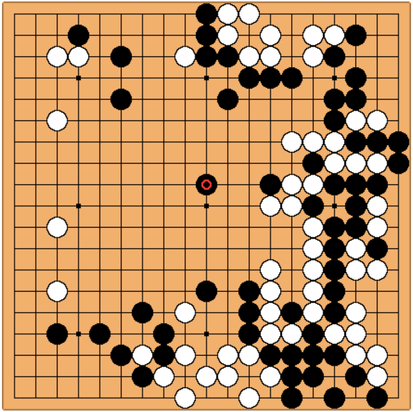
Reinforcement Learning

Chia-Man Hung, Dexiong Chen

Master MVA

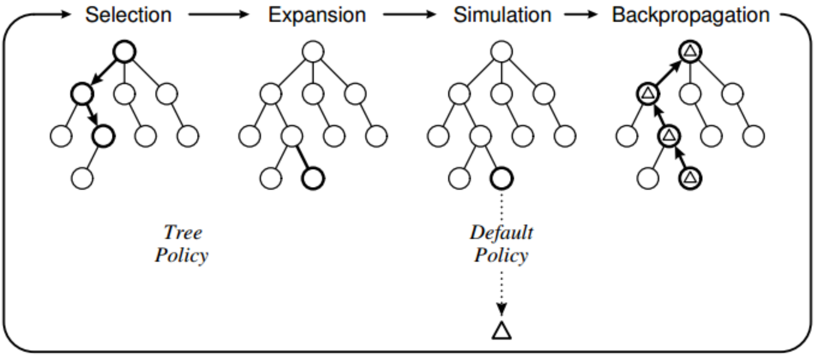
January 23, 2017

- Go: a complex adversarial game
- Infeasibility of the basic MCTS algorithm
- Using a heuristic function can improve performance?



Summary

General Approach



General Approach

Algorithm 1: General MCTS approach.

<u>function</u> MCTSSearch (s_0)
create root node v_0 with state s_0
for *i = 1, ..., itermax* **do**
    $v_l \leftarrow$ TreePolicy(v_0)
    $\Delta \leftarrow$ DefaultPolicy(s(v_l))
    BackPropagate(v_l, Δ)
end

UCT Algorithm

Upper Confidence Bound applied for Trees (UCT)
Tree policy:

$$v^* = \underset{v_c \in \text{child}(v)}{\arg\max} \frac{W(v_c)}{N(v_c)} + K\sqrt{\frac{lnN(v)}{N(v_c)}} \tag{1}$$

where v_c is a child of v, W is the wins count, N is the visits count, and K is a exploration constant to tune.
Exploration vs. Exploitation

Summary

Problem Setting

- Goal: Control a territory
- Influence function:
  The influence function of a white stone (respectively black) at position p over q

$$I_4^W(p,q) = (4 - d_4(p,q))_+, \quad I_4^B(p,q) = -(4 - d_4(p,q))_+, \quad (2)$$

  The total influence of the stones on position q at step t

$$\mathcal{I}_t(q) = \sum_{p \in W_t} I_4^W(p,q) + \sum_{p \in B_t} I_4^B(p,q), \quad (3)$$

- Boundary: Empty, Adversarial

- Reward function:

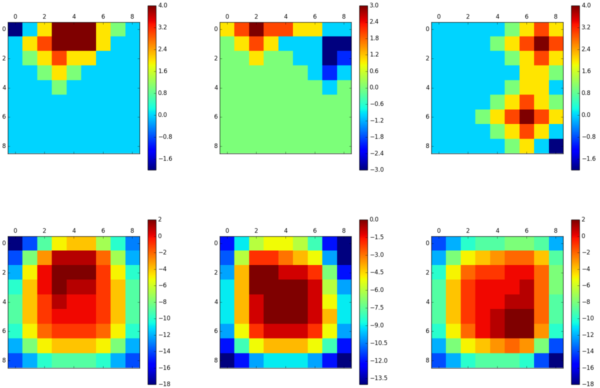
$$r_\tau^W(p) = \sum_{q\in G} (\mathcal{I}_{2\tau}^W(q) - \mathcal{I}_{2\tau-1}^W(q))_+ \mathbb{1}\{\mathcal{I}_{2\tau-1}^W(q) < 0 \leq \mathcal{I}_{2\tau}^W(q)\}$$

$$r_\tau^B(p) = \sum_{q\in G} (-\mathcal{I}_{2\tau+1}^B(q) + \mathcal{I}_{2\tau}^B(q))_+ \mathbb{1}\{\mathcal{I}_{2\tau}^B(q) > 0 \geq \mathcal{I}_{2\tau+1}^B(q)\} \tag{4}$$

The final reward functions for the τ^{th} play of player white (respectively black)

$$\begin{aligned} r_{W,\tau}(p) &= r_\tau^W(p) + c_\tau^W \\ r_{B,\tau}(p) &= r_\tau^B(p) + c_\tau^B. \end{aligned} \tag{5}$$

Illustration of white's reward function



Left: white (1, 3), (0, 5), black (0, 0). Middle: white (0, 2), (0, 6), black (1, 7). Right: white (6, 6), (1, 7), black (8, 8).

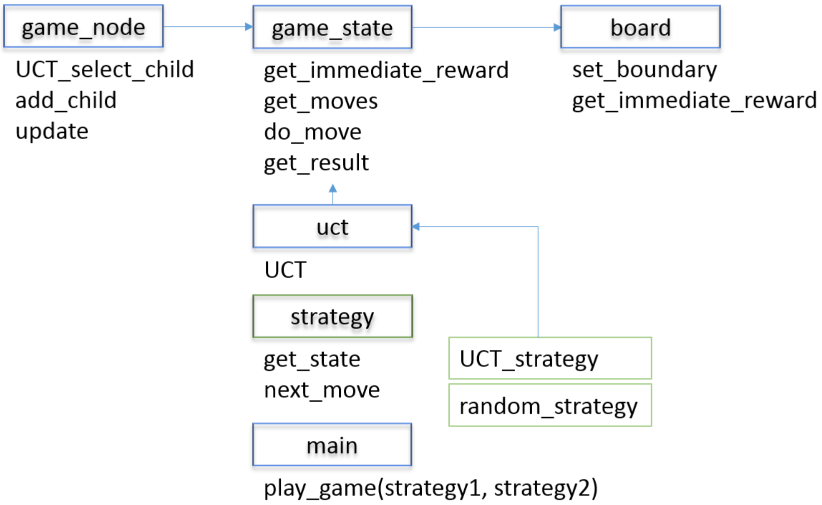
- Pruning: Keep promising children
- Min-Max principle: Take into account the opponent's move

$$a^* = \max_{a \in A(s)} \min_{b \in A(s(a))} r(a, s) - r(b, s(a)) \tag{6}$$

- Back-propagated value: Immediate reward or the official game result (1 win, 0 draw, -1 lose)

Summary

Code Structure



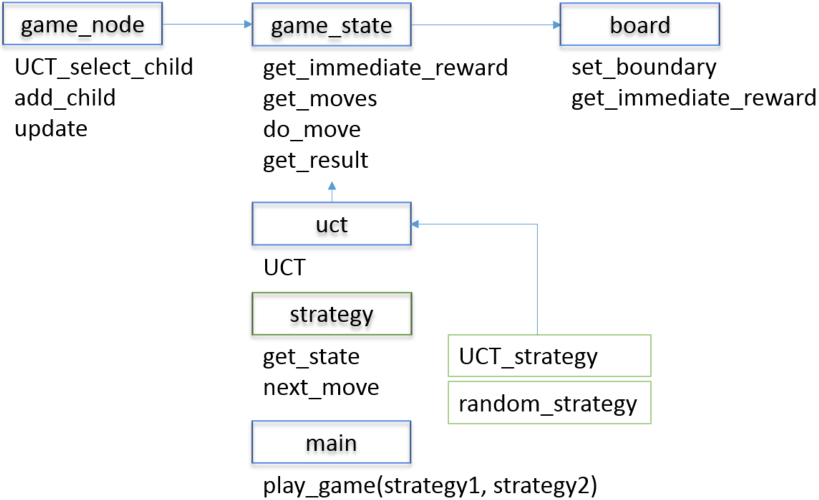
Implementation - UCT

```python
def UCT(rootstate, itermax,  verbose=False):
    """ Conduct a UCT search for itermax iterations starting from rootstate.
        Return the best move from the rootstate.
    """
    rootnode = game_node.GameNode(state=rootstate)

    for i in range(itermax):
        node = rootnode
        state = rootstate.clone()

        # Select
        while node.untried_moves == [] and node.child_nodes != []: # node is fully expanded and non-terminal
            node = node.UCT_select_child()
            state.do_move(node.move)

        # Expand
        if node.untried_moves != []: # if we can expand (i.e. state/node is non-terminal)
            m = random.choice(node.untried_moves)
            state.do_move(m)
            node = node.add_child(m, state) # add child and descend tree
```

Implementation - UCT

```
    # Rollout
    # OpenAI Go board has its maximum limit of moves as 4096
    # state.get_moves() always contains -1
    while not(state.py_pachi_board.is_terminal) and state.nbmoves < 4096 and len(state.get_all_moves()) > 1:
        state.do_move(random.choice(state.get_all_moves()), update=False)

    # Backpropagate
    while node is not None: # backpropagate from the expanded node and work back to the root node
        node.update(state.get_result(node.player_just_moved)) # state is terminal.
        node = node.parent_node
return sorted(rootnode.child_nodes, key = lambda c: c.visits)[-1].move # return the move that was most visited
```

Code Structure



Optimization

In case of non-captures, the influence can be updated easily. This is done in *get_immediate_reward_aux* in *board.py*.

Summary

Which boundary to use? Empty or adversarial?
Compared with the official game result on 1000 games and got similar performance.
$\Rightarrow$ We use the empty boundary in the following.

Scenario 1	
Player A	Random strategy
Player B	UCT strategy: 1000 iterations, without pruning, delayed reward
Wins A/B/draws	2/97/1

The default UCT strategy is better than the random strategy.

Scenario 2	
Player A	UCT strategy: 10 iterations, without pruning, delayed reward
Player B	UCT strategy: 10 iterations, without pruning, immediate reward
Wins A/B/draws	59/40/1

The delayed reward is slightly better than the immediate reward.

Scenario 3	
Player A	UCT strategy: 100 iterations, without pruning, delayed reward
Player B	UCT strategy: 100 iterations, with pruning, ϵ=0, delayed reward
Wins A/B/draws	0/100/0
Scenario 4	
Player A	UCT strategy: 100 iterations, without pruning, immediate reward
Player B	UCT strategy: 100 iterations, with pruning, ϵ=0, immediate reward
Wins A/B/draws	0/100/0

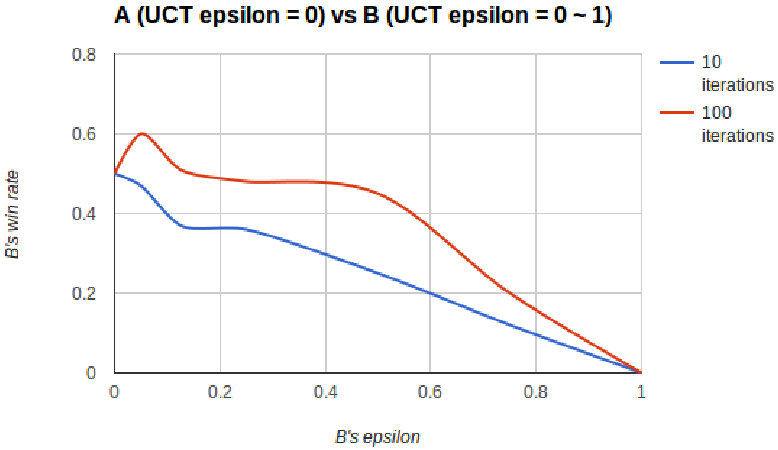
Choosing the optimal action is better than without pruning.

Experiments and Results

Scenario 5	
Player A	UCT strategy: 100 iterations, with pruning, ϵ=0, delayed reward
Player B	UCT strategy: 100 iterations, with pruning, ϵ=0 and min-max, delayed reward
Wins A/B/draws	19/80/1

Considering the min-max principle really boosts the performance.

Scenario 6	
Player A	UCT strategy: 10 iterations, with pruning, ϵ=0, delayed reward
Player B	UCT strategy: 10 iterations, with pruning, ϵ=0.5, delayed reward
Wins A/B/draws	75/25/0
Scenario 7	
Player A	UCT strategy: 100 iterations, with pruning, ϵ=0, delayed reward
Player B	UCT strategy: 100 iterations, with pruning, ϵ=0.5 delayed reward
Wins A/B/draws	55/45/0
Scenario 8	
Player A	UCT strategy: 10 iterations, with pruning, ϵ=0, the delayed reward
Player B	UCT strategy: 10 iterations, with pruning, ϵ=0.25, delayed reward
Wins A/B/draws	64/36/0
Scenario 9	
Player A	UCT strategy: 100 iterations, with pruning, ϵ=0, delayed reward
Player B	UCT strategy: 100 iterations, with pruning, ϵ=0.125, delayed reward
Wins A/B/draws	49/51/0
Scenario 10	
Player A	UCT strategy: 10 iterations, with pruning, ϵ=0, delayed reward
Player B	UCT strategy: 10 iterations, with pruning, ϵ=0.125, delayed reward
Wins A/B/draws	63/37/0


A (UCT epsilon = 0) vs B (UCT epsilon = 0 ~ 1)
B's win rate
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
B's epsilon
10 iterations
100 iterations

Summary

Difficulties

- Simulate the game of Go in the OpenAI Gym.
- From understanding MCTS to actually implementing it. Data structure.
- Experiments are time-consuming (python vs. C++), especially when the min-max principle is considered. (Impossible when min-max level > 2). Ideally, we'd like to have more iterations, otherwise hard to draw conclusion.

Conclusion

- Benefits of the immediate reward: Pruning and the min-max principle boost the performance in general.
- Drawbacks of the immediate reward: The optimal action might be eliminated by pruning. Very slow. Even slower with the min-max principle.
- More iterations will be needed as ϵ grows.
- The choice of a reasonable boundary does not have much influence on the performance.
- Future work: The number of iterations fixed $\rightarrow$ Time budget fixed. (The min-max level can be studied under a fixed time budget.) Optimization with parallel computing. Try other variants combined with the immediate reward.