

MCTS - Connect four

Markov Decision Processes and Reinforcement Learning

Samuele Angheben

Trento University

February 10, 2025

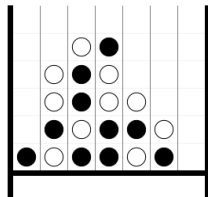
Overview

1. Project Objective
2. Game Overview
3. Implementation
4. MCTS Review
5. Common implementation details
6. Naive MCTS
7. Reinforce Baseline
8. Alphazero
9. Conclusion

The objective of this project is to explore different reinforcement learning algorithm on the game **Connect Four**, starting from very simple mcts implementation to alphazero.

Algorithms:

- **Naive MCTS:** A basic Monte Carlo Tree Search approach without learning.
- **REINFORCE Baseline:** A policy gradient method.
- **AlphaZero:** A self-play RL algorithm combining MCTS with deep learning.



Connect Four is a two-player game played on a vertical 7-column, 6-row grid where players take turns dropping colored discs into the slots, aiming to connect four discs of their color horizontally, vertically, or diagonally.

- **Deterministic:** Outcome fully determined by player actions. $S_{t+1}, R_t = p(S_t, a_t)$
- **Episodic:** Sequence of actions leading to a terminal state. Episode: $\{(S_1, a_1), \dots, (S_T, a_T)\}$
- **Two-Player:** Players 1 and 2 alternate moves, strategies π_1 and π_2 .
- **Zero-Sum:** Rewards sum to zero. $R_1 = -R_2$. Terminal rewards: Win (+1/-1), Draw (0/0)
- **Perfect Information:** Both players have complete knowledge of the current state.
- **Solvability:** Solved. Player 1 can always win with optimal play (Wikipedia). No readily available solvers for testing.

- **Trajectory Length:** Max 42 moves (7 columns x 6 rows).
- **State Space:**
 - Naive: $3^{42} \approx 1.1 \times 10^{20}$
 - Improved: $\approx 4.5 \times 10^{12}$ states
 - Storage: 4.5×10^{12} states \times 8 bytes/state = 36 TB (64 bit state representation)
- **Tabular Methods:** Infeasible due to large state space.

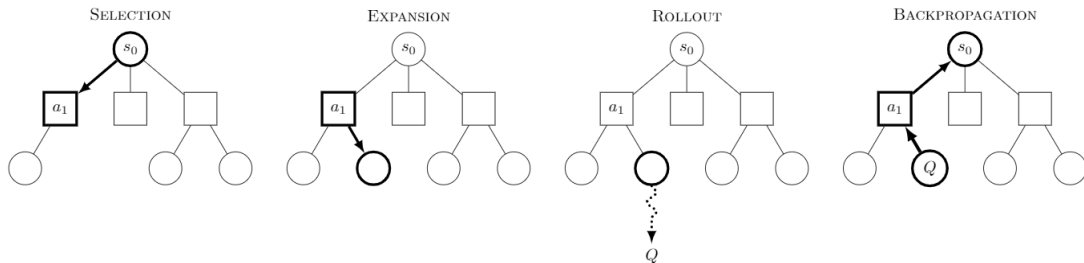
Framework: JAX provides a high-performance framework for computational efficiency and flexibility, allowing for scalable and efficient development of machine learning algorithm.

JAX Features:

- JIT Compilation
- GPU/TPU Acceleration
- Efficient Vectorization (vmap)
- Automatic Parallelization (pmap)
- Functional Programming (Pure Functions)

- `mctx`: JAX-native MCTS implementations.
- `pgx`: Game environment simulation, diverse games, fast parallelization.
- `Haiku`: Neural network building and training in JAX.
- `optax`: Gradient processing and optimization.

Monte Carlo Tree Search is a decision-making algorithm that explores a decision space by constructing a search tree through repeated simulations. Given a computation budget it balances exploration and exploitation to refine the root node policy.



To perform a MCTS search using the MCTX library, two main functions must be provided:

RootFnOutput(state):

- Specifies the representation of the root state.
- Returns the prior logits and the estimated value of the root state.

recurrent_fn(state, action):

- Encapsulates the environment dynamics.
- Returns the reward, discount factor, and for the new state, the prior logits and value.

The search returns **action_weight**, representing the updated root node policy (policy_out).

In alternating-turn games, a negative discount factor ($\gamma = -1$) is used to invert value estimates between players:

$$V(S_t) = \mathbb{E}[r_{t+1} + \gamma \cdot V(S_{t+1})] = \mathbb{E}[r_{t+1} + (-1) \cdot V(S_{t+1})]$$

- S_t : Current state (Player 1's turn).
- S_{t+1} : Next state (Player 2's turn), where the value $V(S_{t+1})$ represents the opponent's optimal value.

This inversion maintains the zero-sum property by ensuring that a high value for one player corresponds to a low value for the opponent.

The neural network is a custom **ResNet**, a convolutional network with residual connections and batch normalization. It has two output heads:

$$l, v' = \text{ResNet}_\theta(S_t)$$

1. Policy Head: Produces a probability distribution over possible actions:

$$\pi(a \mid S_t) \approx \hat{\pi}(a \mid S_t) = \text{softmax}(l)$$

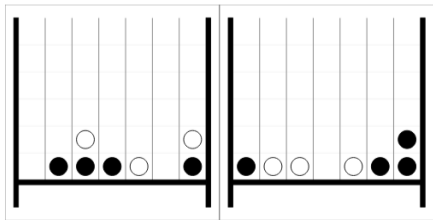
2. Value Head: Outputs a scalar value approximating the value function:

$$V(S_t) \approx \hat{V}(S_t) = v'$$

Naive implementation of Monte Carlo Tree Search:

- **no learning**
- MCTS
 - policy: uniform
 - value: single random rollout

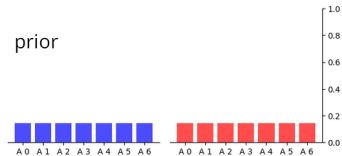
```
61
62 def RecurrentFn(s, a)
63     s', reward <- env_step(s, a)
64     V, pi <- random_rollout(s'), uniform_prior
65     discount <- -1.0
66     return s', pi, V, reward, discount
67
68 def RunMCTS(s)
69     root = (V, pi(a|s)) <- random_rollout(s), uniform_prior
70     policy_output <- MCTS(root, RecurrentFn, num simulation)
71     return arg max_a policy_output
72
```



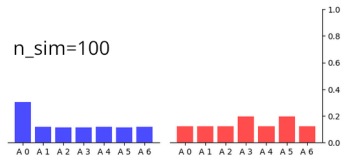
Value = [-1, -1]

Policy output

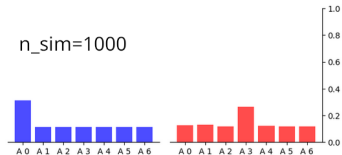
prior



n_sim=100



n_sim=1000



Value of the state: [-1, -1.]

Reinforce baseline algorithm:

- Policy based algorithm
- no MCTS
- on policy, batch
- neural network function and policy approximation


```

1 Initialize: Neural Network, Optimizer, Replay Buffer
2
3 Main Loop:
4   for iteration in 1..MaxIterations:
5     # Self-play Phase
6     Collect Games:
7       for N games:
8         def step_fn(s):
9           _, pi_nn(a|s) <- nn_forward(s)
10          a ~ pi_nn(a|s)
11
12          s', r <- env_step(s, a)
13
14          return (s, a, s', r, discount=-1),
15                s'
16
17   s_init <- env.reset()
18   data <- jax.scan(step_fn, s_init)
19
20   samples <- compute_loss_input(data) # cumulative rewards
21   Save samples to Replay Buffer # (s_t, a_t, G_t:value target)
22
23
24   Shuffle samples and make minibatches (Replay Buffer)
25
26

```

```

27
28
29
30 # Training Phase
31 Train Network:
32   batch ~ Replay Buffer # sample (s_t, a_t, G_t)
33   compute_loss:
34     V_nn, pi_nn(a|s_t) <- nn_forward(s_t)
35     A(s_t, a_t) <- G_t - V_nn(s_t) # compute advantage
36     policy_loss <- -mean[log(pi_nn(a_t|s_t) * A(s_t, a_t))]
37     value_loss <- MSE(V_nn(s_t), G_t)
38     entropy_loss <- entropy(pi_nn)
39
40     loss <- policy_loss + value_loss + 0.1 * entropy_loss
41
42   Compute gradient (loss)
43   Update model parameters
44
45 # Saving Phase (periodic)
46 if iteration % saving_interval == 0:
47   Save checkpoint
48
49
50
51
52

```

Total Loss:

$$L = \mathcal{L}_{\text{policy}} + \mathcal{L}_{\text{value}} + 0.1 \times \mathcal{L}_{\text{entropy}}$$

Where:

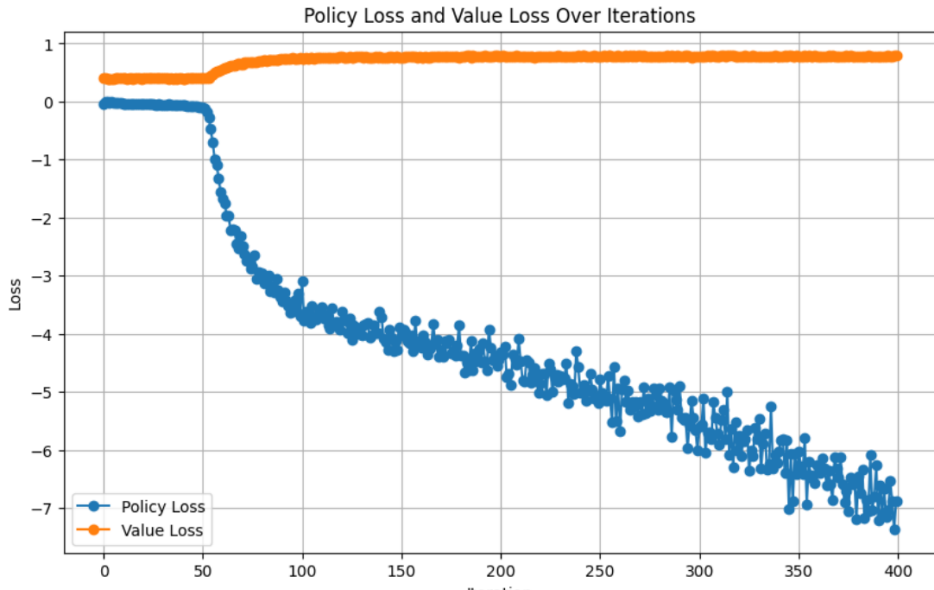
$$\mathcal{L}_{\text{policy}} = -\mathbb{E} [\log \hat{\pi}(a_t | S_t) \cdot A(S_t, a_t)]$$

$$A(S_t, a_t) = G_t - \hat{V}(S_t)$$

$$\mathcal{L}_{\text{value}} = \mathbb{E} \left[\left(\hat{V}(S_t) - G_t \right)^2 \right]$$

$$\mathcal{L}_{\text{entropy}} = -\mathbb{E} \left[\sum_a \pi(a | S_t) \log \pi(a | S_t) \right]$$

Gradient Clipping: ensures that the gradients do not exceed a specified threshold, preventing exploding gradients.



Even with loss improvements, divergence still occurs.

The Deadly Triad:

Bootstrapping ×

Function Approximation ✓

Off-Policy ×

Alphazero algorithm

- MCTS
 - policy: neural network policy head
 - value: neural network value head

```

1 Initialize: Neural Network, Optimizer, Replay Buffer
2
3 Main Loop:
4   for iteration in 1..MaxIterations:
5     # Self-play Phase
6     Collect Games:
7       for N games:
8         def step_fn(s):
9           root = V_nn, pi_nn(a|s) <- nn_forward(s)
10          run MCTS(root, num_simulation):
11            s', reward <- env
12            V', pi <- nn_forward
13
14          a <- MCTS.policy_out.action
15
16          s', r <- env_step(s, a)
17
18          return (s, policy_out, s', r, discount=1),
19                s'
20
21      s_init <- env.reset()
22      data <- jax.scan(step_fn, s_init)
23
24      samples <- compute_loss_input(data) # cumulative rewards
25      Save samples to Replay Buffer # (s_t, policy_out, G_t:value target)
26
27      Shuffle samples and make minibatches (Replay Buffer)
28
35
36
37
38
39 # Training Phase
40 Train Network:
41   batch ~ Replay Buffer # sample (s_t, policy_out_t, G_t)
42   compute_loss:
43     V_nn, pi_nn(a|s_t) <- nn_forward(s_t)
44     policy_loss <- cross_entropy(pi_nn(a|s_t), policy_out)
45     value_loss <- MSE(V_nn(s_t), G_t)
46
47     loss <- policy_loss + value_loss
48
49   Compute gradient (loss)
50   Update model parameters
51
52 # Saving Phase
53 if iteration % saving_interval == 0:
54   Save checkpoint
55
56
57
58
59
60
61
62

```

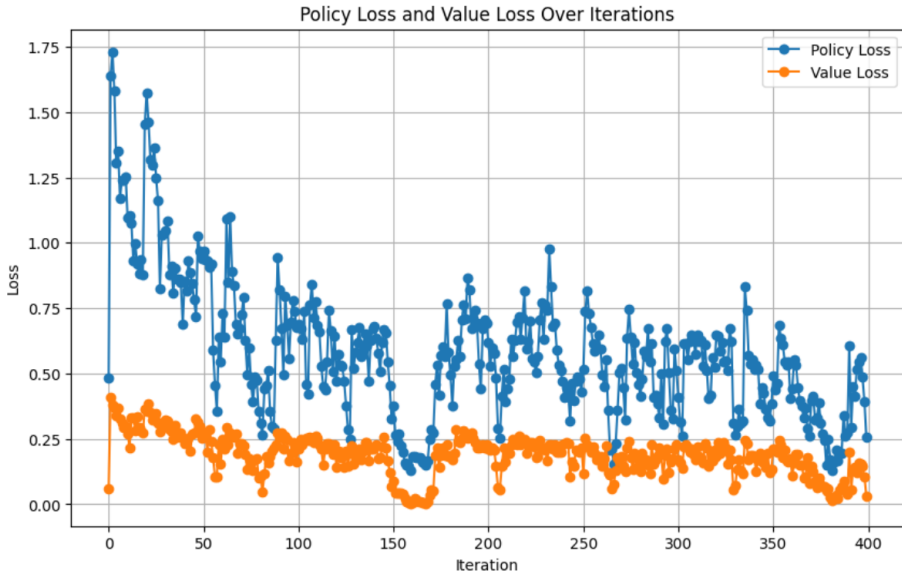
Total Loss:

$$L = \mathcal{L}_{\text{policy}} + \mathcal{L}_{\text{value}}$$

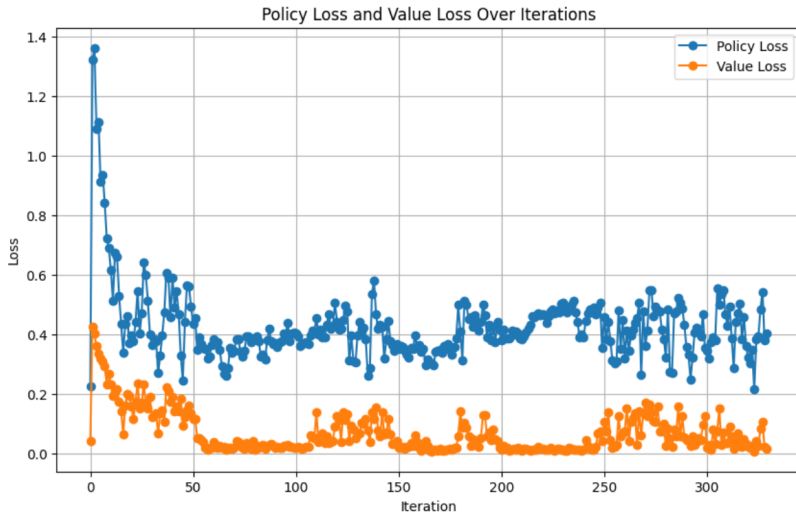
Where:

$$\mathcal{L}_{\text{policy}} = -\mathbb{E}_{a \sim \text{policy_out}} [\log \hat{\pi}(a)]$$

$$\mathcal{L}_{\text{value}} = \mathbb{E} \left[\left(\hat{V}(S_t) - G_t \right)^2 \right]$$

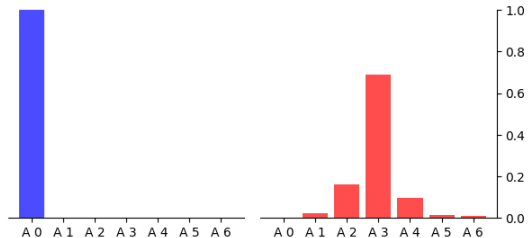
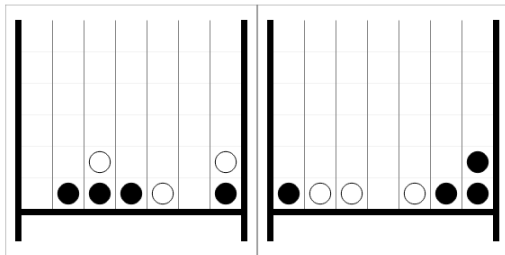


With better training: increase num. MCTS simulation, batch size, neural network



```
61
62 def RecurrentFn(s, a)
63     s', reward <- env_step(s, a)
64     V, pi <- nn_forward(s')
65     discount <- -1.0
66     return s', pi, V, reward, discount
67
68 def RunMCTS(s)
69     root = (V, pi(a|s)) <- nn_forward(s)
70     policy_output <- MCTS(root, RecurrentFn, num simulation)
71     return arg max_a policy_output
72
```

```
74
75 def RunOneShot(s)
76     pi_nn(a|s) <- nn_forward(s)
77     return arg max_a pi_nn(a|s)
```



Array([-0.93581444, 0.85516936], dtype=float32)

Notebook games and live play

- hyper-parameter tuning for better training
- longer training
- network architecture
 - increase capacity
 - transformer architecture
- MCTS: test different selection algorithm

References



Danihelka, I., Guez, A., Schrittwieser, J., and Silver, D. (2022).

Policy improvement by planning with gumbel.

<https://www.openreview.net/forum?id=bERaNdоеgn0>.

Accessed: 2025-02-07.



Koyamada, S., Okano, S., Nishimori, S., Murata, Y., Habara, K., Kita, H., and Ishii, S. (2024).

Pgx: Hardware-accelerated parallel game simulators for reinforcement learning.



Silver, D., Huang, A., Maddison, C. J., and et al. (2016).

Mastering the game of go with deep neural networks and tree search.

Nature, 529(7587):484–489.



Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2017a).

Mastering chess and shogi by self-play with a general reinforcement learning algorithm.



Silver, D., Schrittwieser, J., Simonyan, K., and et al. (2017b).

Mastering the game of go without human knowledge.

Nature, 550:354–359.

Thank you for your attention!