Fraunhofer IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

Reinforcement Learning

—

Exercise 9: MCTS

16.06.2023

Sebastian Rietsch

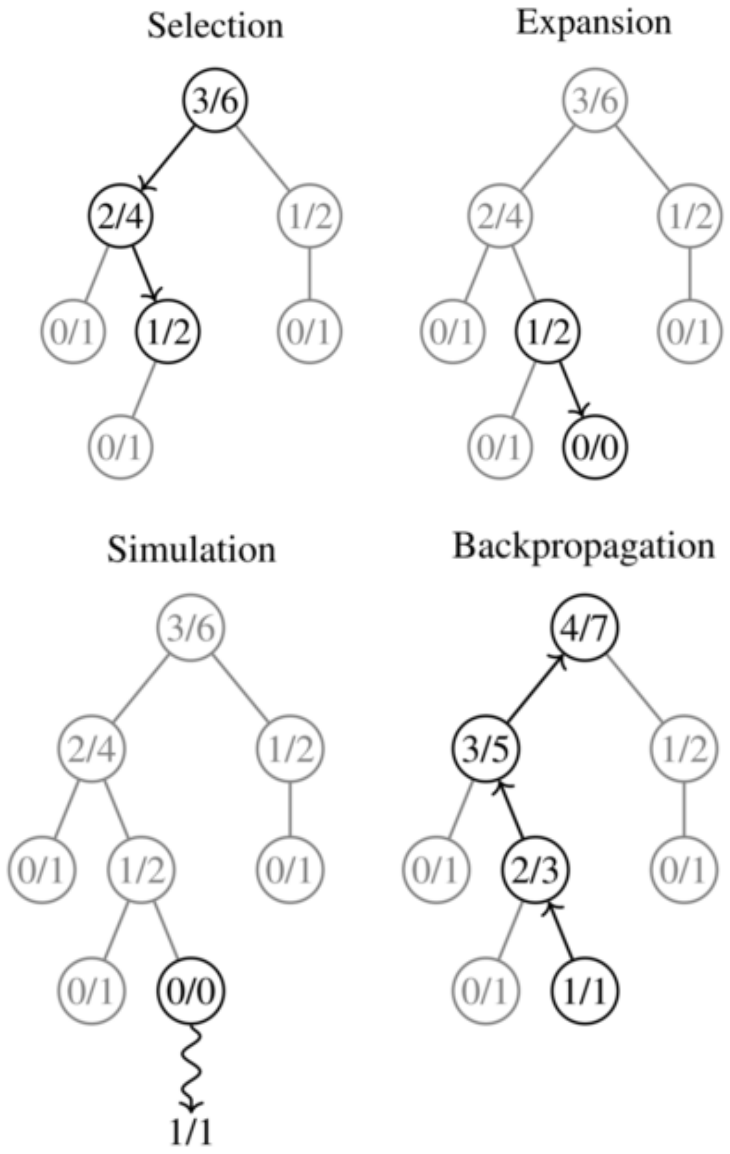
Exercise Sheet 7

Policy Gradient

Fraunhofer IIS

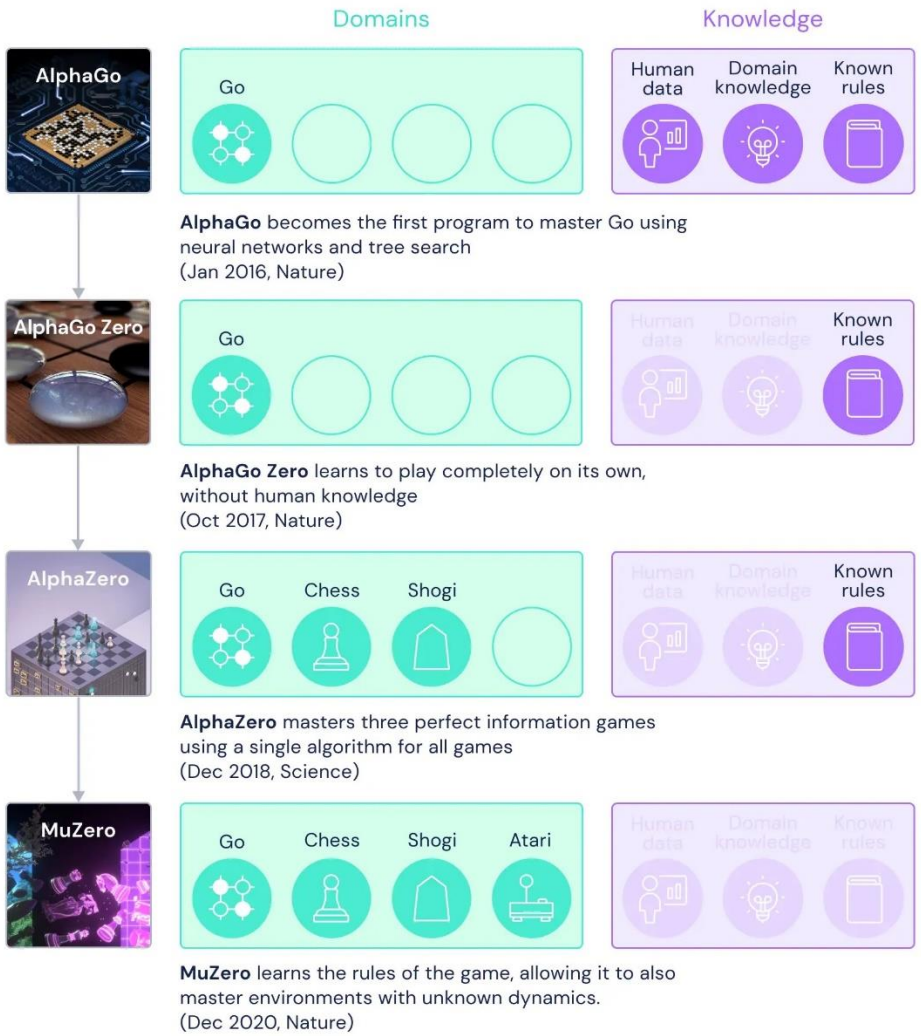
Monte Carlo Tree Search

- Heuristic search algorithm using random sampling for (deterministic) problems
  - In our setting: Nodes are states, edges are actions
- Play many rollouts from the root node
  - **Selection**: Select successive child nodes until a leaf node is reached
  - **Expansion**: Create a new child node
  - **Simulation:** Continue with (random) actions until the terminal state
  - **Backpropagation:** Update information in the nodes on the path traversed
- Balancing exploitation and exploration during expansion via **UCT** formula
  - $a = argmax_i \frac{w_i}{n_i} + c\sqrt{\frac{\ln N_i}{n_i}}$



Fraunhofer IIS

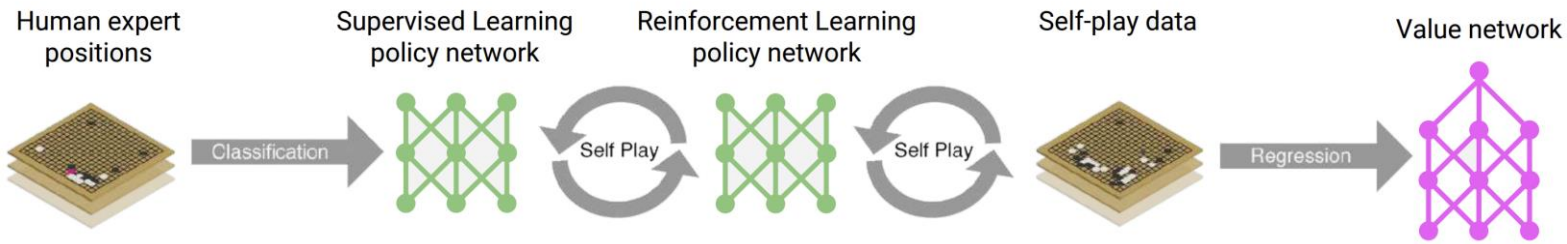
The Evolution of AlphaGo to muZero



https://www.deepmind.com/blog/muzero-mastering-go-chess-shogi-and-atari-without-rules

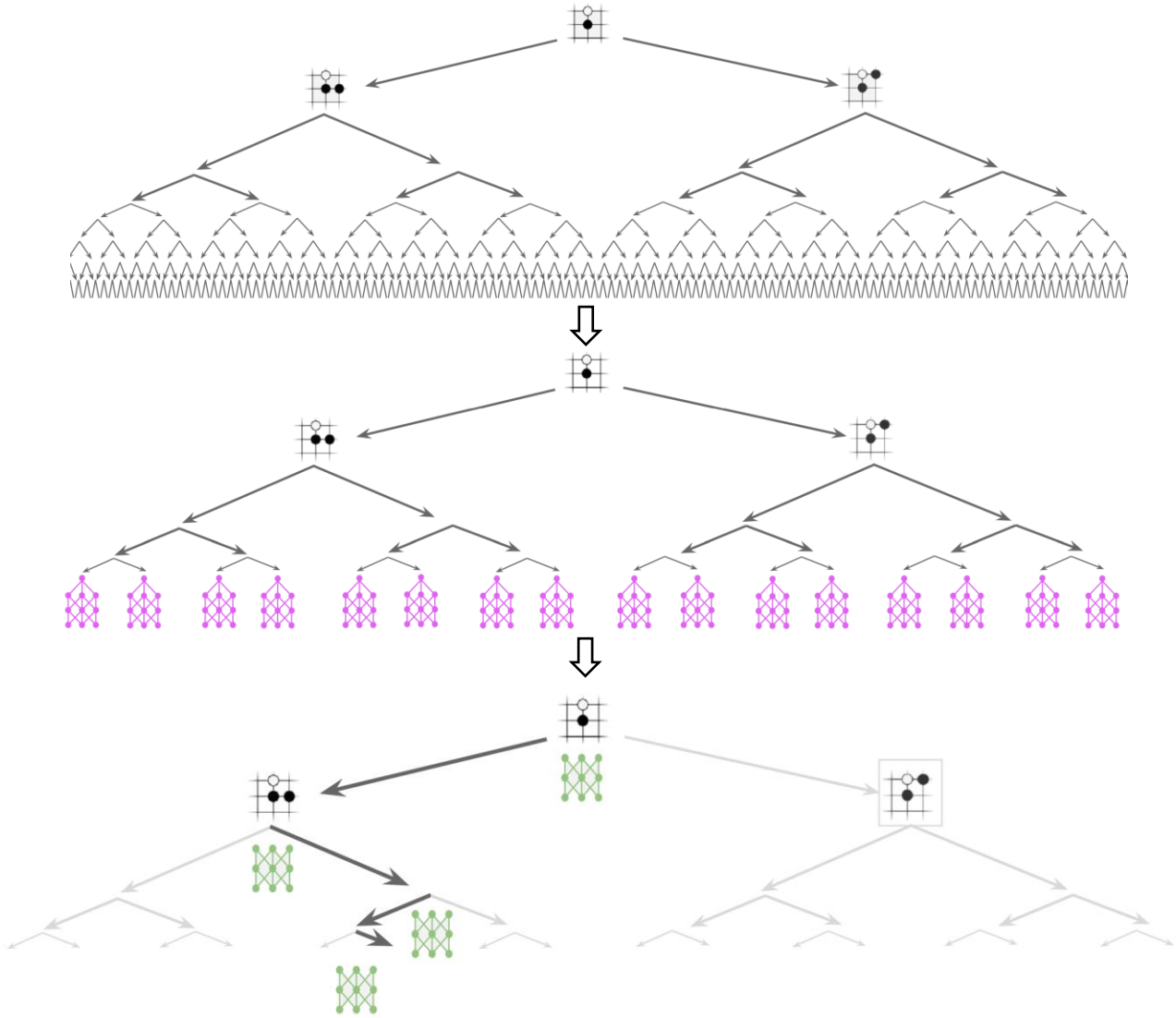
AlphaGo

- **AlphaGo** defeated the Go champion Lee Sedol in a best-of-five tournament in 2016
- Algorithm outline
  - **Training**
    - A policy $p(s|a)$ is trained to predict human expert moves in a data set of positions, refined via policy gradient through self-play, and training of value regressor on self-play data
  - **Deployment**
    - MCTS with policy and value network



Fraunhofer IIS

AlphaGo – Influences on Search Complexity



Exhaustive search

Reducing depth with value network

Reducing breath with policy network

https://www.davidsilver.uk/wp-content/uploads/2020/03/AlphaGo-tutorial-slides_compressed.pdf

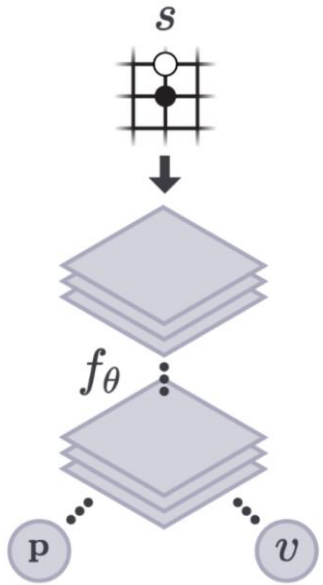
Fraunhofer IIS

AlphaZero

- One deep neural network $f_\theta(s) = (p, v)$ with
  - move probabilities $p = \Pr(a|s)$ and
  - value prediction v (win probability of the current player)
- *"Tabula rasa"* reinforcement learning
  - A policy plays against a past version of itself (self-play)
  - In each position, an MCTS search is executed
    - Guided by the neural network's move probabilities p
    - More robust, sophisticated policy (tree-search informed by policy network's "best guess")
  - Network is updated towards MCTS move probabilities (policy head) and self-play winner outcome (value head)

"policy evaluation"

"policy improvement"



  - "Policy iteration procedure"

Fraunhofer IIS

Fraunhofer
IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

Thank you for your attention!