Stochastic Planning in Games: an AlphaGo Case-study

Tommaso Macri
AlphaGo beats 18-time world champion Lee Sedol 4 games to 1

The New York Times
It isn’t looking good for humanity.

In a major breakthrough for artificial intelligence, AlphaGo Zero took just three days to master the ancient Chinese board game of Go ... with no human help
The Game of Go
The Game of Go, Rules

Aim of the game is to surround more territory than the opponent
Comparison with Chess

- ~250 legal moves per position
- ~150 moves per game

- ~35 legal moves per position
- ~80 moves per game
Artificial Intelligence perspective

Go game has challenged artificial intelligence researchers for many decades

A Go board configuration
Mastering the Game of Go: a Major Breakthrough for AI

David Silver¹*, Aja Huang¹*, Chris J. Maddison¹, Arthur Guez¹, Laurent Sifre¹, George van den Driessche¹, Julian Schrittwieser¹, Ioannis Antonoglou¹, Veda Panneershelvam¹, Marc Lanctot¹, Sander Dieleman¹, Dominik Grewe¹, John Nham², Nal Kalchbrenner¹, Ilya Sutskever², Timothy Lillicrap³, Madeleine Leach³, Koray Kavukcuoglu¹, Thore Graepel¹ & Demis Hassabis¹
Planning and RL
Planning vs Learning

In Planning, we use the simulated experience to update the value function and policy.

In Learning we use Experience Generated by the Environment (not simulated).
# Planning vs Learning: the Dyna-Q example

## Direct RL

1. Initialize $Q(s, a)$ and $Model(s, a)$ for all $s \in S$ and $a \in A(s)$
2. Loop forever:
   a. $S \leftarrow$ current (nonterminal) state
   b. $A \leftarrow \varepsilon$-greedy($S, Q$)
   c. Take action $A$; observe resultant reward, $R$, and state, $S'$
   d. $Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$
   e. $Model(S, A) \leftarrow R, S'$ (assuming deterministic environment)
   f. Loop repeat $n$ times:
      - $S \leftarrow$ random previously observed state
      - $A \leftarrow$ random action previously taken in $S$
      - $R, S' \leftarrow Model(S, A)$
      - $Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$

## Planning

- Reinforcement Learning, Sutton et al.
Planning vs Learning: pros and cons

Direct vs Undirect Learning

Policy/value functions

Environment

Model

simulated experience
search control

real experience
model learning
direct RL update
planning update

Reinforcement Learning, Sutton et al.
MCTS
(Monte Carlo Tree Search)
Monte Carlo Tree Search

Repeat while time remains

Selection → Expansion → Simulation → Backup

Tree Policy

Rollout Policy

Reinforcement Learning, Sutton et al.
Monte Carlo Tree Search

Selection
Expansion
Simulation
Backpropagation

https://en.wikipedia.org/wiki/Monte_Carlo_tree_search
The AlphaGo Breakthrough
Mastering the game of Go with deep neural networks and tree search

David Silver¹, Aja Huang², Chris J. Maddison³, Arthur Guez⁴, Laurent Sifre⁵, George van den Driessche⁶,
Julian Schrittwieser⁷, Ioannis Antonoglou⁸, Veda Panneershelvam⁹, Marc Lanctot¹⁰, Sander Dieleman¹¹, Dominik Grewe¹²,
John Nham¹³, Nal Kalchbrenner¹⁴, Ilya Sutskever¹⁵, Timothy Lillicrap¹⁶, Madeleine Leach¹⁷, Koray Kavukcuoglu¹⁸,
Thore Graepel¹⁹ & Demis Hassabis¹
AlphaGo

- Go is a perfect information game.
- Compute optimal value function?

- Tree search = $b^d$
  - $b = 250$; $d = 150$

- Exhaustive search is infeasible
AlphaGo: the 4 Neural Networks

- **Rollout policy** ($p_\pi(a|s)$)
  - Forward Propagation: 2µs
  - Accuracy: 24%

- **SL policy network** ($p_\sigma(a|s)$)
  - Forward Propagation: 3ms
  - Accuracy: 57%

- **RL policy network** ($p_\rho$)
  - Policy gradient

- **Value network** ($v_\theta$)
  - Self-Play
  - Regression

**Data**
- Human expert positions
- Self-play positions

Mastering the game of go with deep neural networks and tree search, David Silver et al.
AlphaGo: multiple outputs for the policy and single for the value

Mastering the game of go with deep neural networks and tree search, David Silver et al.
AlphaGo: combining Policy Network and Value Network with MCTS

Mastering the game of go with deep neural networks and tree search, David Silver et al.
Mastering the game of Go without human knowledge

David Silver\textsuperscript{1,}*, Julian Schrittwieser\textsuperscript{1,}*, Karen Simonyan\textsuperscript{1,}*, Ioannis Antonoglou\textsuperscript{1}, Aja Huang\textsuperscript{1}, Arthur Guez\textsuperscript{1},
Thomas Hubert\textsuperscript{1}, Lucas Baker\textsuperscript{1}, Matthew Lai\textsuperscript{1}, Adrian Bolton\textsuperscript{1}, Yutian Chen\textsuperscript{1}, Timothy Lillicrap\textsuperscript{1}, Fan Hu\textsuperscript{1}, Laurent Sifre\textsuperscript{1},
George van den Driessche\textsuperscript{1}, Thore Graepel\textsuperscript{1} & Demis Hassabis\textsuperscript{1}
AlphaGo Zero: Only one Neural Network for Policy and Value

\[(p, v) = f_\theta(s)\]
AlphaGo Zero: Using MCTS to select moves throughout self-play

Mastering the game of go without human knowledge, David Silver et al.
A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play

David Silver\textsuperscript{1,2,}\dagger, Thomas Hubert\textsuperscript{1*}, Julian Schrittwieser\textsuperscript{1*}, Ioannis Antonoglou\textsuperscript{1}, Matthew Lai\textsuperscript{1}, Arthur Guez\textsuperscript{1}, Marc Lanctot\textsuperscript{1}, Laurent Sifre\textsuperscript{1}, Dharshan Kumaran\textsuperscript{1}, Thore Graepel\textsuperscript{1}, Timothy Lillicrap\textsuperscript{1}, Karen Simonyan\textsuperscript{1}, Demis Hassabis\textsuperscript{1,}
Alpha Zero: Learning and MCTS do not assume symmetry

Alpha Go and Alpha Go Zero assumed symmetry for:
• Training data augmentation
• Bias removal in Monte Carlo evaluations

Alpha Zero considers also the “Drawn” outcome

Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, David Silver et al.
Alpha Zero: The Algorithm Architecture

Alpha Go Zero, Alpha Zero:
• one Neural Network (Policy + Value)
• Monte Carlo Tree Search

Alpha Go Zero
• Wait for an iteration to conclude to update NN
• Compare the new policy to the best

Alpha Zero
• Update the NN continuously
• Always generate Self-Play with the latest NN
Results: AlphaGo beats the European Go Champion

Mastering the game of go with deep neural networks and tree search, David Silver et al.
AlphaGo Zero: better performances than AlphaGo

Mastering the game of go without human knowledge, David Silver et al.
AlphaGo Zero: Learns human expert moves and beyond

Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, David Silver et al.
Alpha Zero: program applied to Chess and Shogi

Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, David Silver et al.
Conclusions

I can’t disguise my satisfaction that it plays with a very dynamic style, much like my own!"

GARRY KASPAROV
FORMER WORLD CHESS CHAMPION

The implications go far beyond my beloved chessboard... Not only do these self-taught expert machines perform incredibly well, but we can actually learn from the new knowledge they produce."

GARRY KASPAROV
FORMER WORLD CHESS CHAMPION

Tommaso Macrì

Stochastic Planning in Games: an AlphaGo Case-study