AlphaPokerZero?
Deep Reinforcement Learning for Imperfect Information Games

Edgard Bonilla, Divyanshu Murli, Geoffrey Penington

edgard@stanford.edu, divyansh@stanford.edu, geoffp@stanford.edu,

Introduction

- Deep reinforcement learning has achieved superhuman, state-of-the-art performance in many perfect information games (e.g. Chess, Go, Shogi), but imperfect information games such as Poker are still dominated by more traditional reinforcement learning techniques (CFR).
- Basic deep reinforcement learning algorithms (DQN, AlphaZero etc.) fail for imperfect information games.
- Heinrich and Silver (2017) developed Neural Fictitious Self Play (NFSP). This combines a DQN with an average policy network to achieve stable learning.
- Our goal: adapt the AlphaZero algorithm to use in imperfect information games and apply it to versions of poker.

The algorithm

- Single policy-value network takes the players information set as an input and returns a policy p together with an estimate v of the value of the position.
- Opponent range network also takes the players information set as its input and return an estimate of the opponent cards.
- A Monte Carlo Tree Search (MCTS) is used to estimate a best response strategy to the current policy p. The player uses an upper confidence bound (UCB) strategy guided by the policy p and value estimate v, while the opponent uses Monte Carlo sampling of the policy p given cards sampled using the opponent range network.
- Games are mostly played using the policy p, with only a small fraction ($\eta = 0.1$) using the best response found via the tree search (anticipatory learning).
- The policy p is trained to approximate the average of all best response policies from the entirety of training. Reservoir sampling is used when the data exceeds the memory buffer.
- The value estimate v and opponent range estimate are trained using a smaller dataset of all recent games (regardless of which policy was used) to approximate the actual outcomes of games and opponent cards respectively.
- Cross entropy loss used for policy and opponent range. L2 loss used for value estimate. L2 regularisation.
- The temperature of the best response policy ($N^{1/\tau}(s,a)$ where $N(s,a)$ is the number of times action a was taken in information set s by the tree search) was gradually decreased over time to speed up learning. To avoid biasing the best response policy in favour of action with higher policy probability, the root node initial policy was replaced by p^τ.

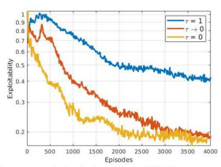
Leduc Poker

- We used a simplified version of poker known as Leduc as a testbed to measure performance. This made it possible to exactly calculate the exploitability of the strategies we produced.
- Our algorithm successfully converged to an approximate Nash equilibrium. The lowest exploitability achieved (~ 0.16) was roughly twice the best exploitability reached by Heinrich et al. using NFSP.
- Experience with Leduc helped optimise hyperparameters which could then be used for Texas Hold Em.
- Each network used two fully connected layers. The memory sizes for the short term (100k) and reservoir (1M) memories were taken from NFSP.

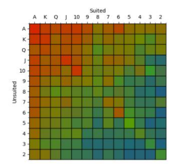
Heads Up Limit Texas Hold Em

- We also tested our algorithm on Heads Up Limit Texas Hold Em, a 'real' poker game whose Nash equilibrium was only found in 2015 using a supercomputer.
- Because each player has two cards, the opponent range network was adapted so that we could first sample a single card, and then feed this card back into the network in order to sample a second card, as in language sampling models.
- The model beat simple strategies and produced strategies that agreed with human intuition.
- Performance of best response strategy against policy p appeared to be decreasing over time, although it is hard to distinguish this effect from noise. This suggests the policy is becoming harder to exploit.

Results

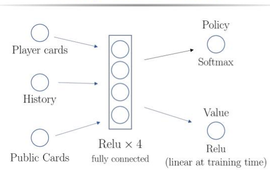


Performance at Leduc Poker with the MCTS temperature parameter set to $\tau = 1$ (blue), $\tau = 0$ (yellow) and gradually decaying from $\tau = 1$ (red). Zero exploitability corresponds to the Nash equilibrium solution.

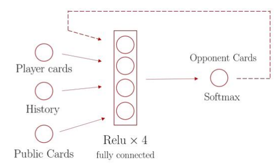


Preflop strategy for Heads Up Limit Texas Hold Em after 2000 iterations. Red indicates raise, green call and blue fold. The upper right triangle shows the strategy for suited cards, the lower left triangle shows the strategy for unsuited cards.

Policy-Value Network



Opponent Range Network



Discussion

- Our algorithm successfully converged to an approximate Nash equilibrium in Leduc Poker. It provides an alternative approach to NFSP for deep reinforcement learning in imperfect information games.
- The performance at Leduc Poker was not quite as high as for NFSP. However, as we have optimised hyperparameters and algorithmic details, the performance of our algorithm has steadily improved. We are confident that further performance gains are possible.
- MCTS algorithms are most effective for very deep games. This suggests our algorithm should be even more useful in 'real poker', and especially in even deeper imperfect information games such as various board games.
- Our algorithm appears to be learning Heads Up Limit Texas Hold Em successfully but we were limited in our ability to measure performance, because of the time required to simulate sufficient games given our limited computational power.

Future

- Future plans include continuing to optimise performance on Leduc poker. Just in the last few days, various optimisations have improved the exploitability achieved by a factor of two.
- We will also look to test our algorithm's performance at Limit Texas Hold Em against state-of-the-art poker bots.
- We can apply our algorithm to Heads Up No Limit Texas Hold Em, where superhuman performance has only been achieved in the last year, as well as poker games with more than two players, where humans still reign supreme.
- Finally, the algorithm can be easily adapted to various other imperfect information games, particularly games where very deep tree searches are vital to success.

References

- D. Silver et al. *Mastering the game of go without human knowledge*, Nature 550.7676 (2017): 354
- D. Silver et al,*Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*, arXiv:1712.01815
- M. Moravcik et al, *Deepstack: Expert Level Artificial Intelligence in No-Limit Poker*, arXiv:1701.01724
- J. Heinrich and D. Silver, Deep *Reinforcement Learning from Self-Play in Imperfect Information Games*, arXiv:1603.01121