Machine Learning Project: Report 2

Ignace Bleukx Quinten Bruynseraede

May 3, 2020

1 Introduction

1.1 Evaluation metrics

Evaluation of agents is traditionally done using **NashConv** and **exploitability**. We introduce these concepts here, using terminology consistent with Lanctot et al. [1].

Given a two-player policyπ, we say that π_i^b is the best response for player i. The best response is defined as the policy that maximizes payoff for player i, given the policies of other players (pi_{-i}).

We then define the incentive to change policies $d_i(\pi)$ as $d_i(\pi) = u_i(\pi_i^b, \pi_{-i}) - u_i(\pi)$, i.e. the possible gain in value when switching to a best-response strategy. It is clear that this can be used as a metric to evaluate a policy: a Nash equilibrium is found when $d_i(\pi) = 0$ (the policy cannot be improved given other player's policies). Any value > 0 captures room for improvement for a given theory.

From this initial notion of policy evaluation, the **NashConv** metric is derived as the sum of $d_i(\pi)$ for both players.

Another metric that is often used, is **exploitability**. For two-player zero-sum games (such as the games we will be learning in this assignment), exploitability equals $\frac{NashConv}{2}$. We therefore only use Exploitability to evaluate our policies (as NashConv can be derived from Exploitability in this case).

Looking for Nash equilibria is interesting in zero-sum games because they guarantee a maximal payoff against any policy the other players might have. We will therefore focus on finding approximations of Nash equilibria (i.e. minimizing exploitability). It should also be noted that convergence toward Nash equilibria is a property of individual algorithms, and is not guaranteed.

1.2 Algorithm 1: Fictitious Self-Play

1.2.1 Extension: Neural Fictitious Self-Play

1.3 Algorithm 2: Counterfactual Regret Minimization

Counterfactual Regret Minimization (CFR in short) is an algorithm that is designed to find Nash equilibria in large games. It introduces the notion of counterfactual regret, and minimizes this to compute a Nash Equilibrium. Zinkevich et al. [2] show that CFR can solve games with up to 10^{12} states, such as Texas Hold'em. We therefore expect it to perform well on reduced variants of traditional poker, such as Kuhn and Leduc poker.

An important concept in reinforcement learning is the notion of regret, which can be paraphrased as the difference between maximum and actual payoff an agent receives when executing a sequence of steps. The goal of many reinforcement learning algorithms is to minimize regret. What makes regret minimization less applicable in large games,

is that regret is to be minimized over all game states. This quickly becomes infeasible when dealing with a large number of states. The general idea of counterfactual regret minimization is to decompose regret into a set of regret terms, whose som approaches the actual regret. Zinkevich et al. then show that individually minimizing these regret terms leads to a Nash equilibrium. We will now introduce counterfactual regret minimization more formally.

- A history h is a sequence of actions performed by the agents, starting from the root of the game tree (the initial state).
- An information state I consists of a player and the information that is visible to that player.
- All player have their individual strategies, and we call the combination of these strategies at time t: σ^t.

1.3.1 Extension: Regression Counterfactual Regret Minimization

1.3.2 Extension: Counterfactual Regret Minimization against best responder

1.3.3 Extension: Deep Counterfactual Regret Minimization

2 Kuhn Poker

- Which algorithm is most suitable to develop an agent to play Kuhn Poker, minimizing exploitability?
- Can we exploit properties of Kuhn Poker to optimize parameters?

3 Leduc Poker

- Which algorithm is most suitable to develop an agent to play Leduc Poker, minimizing exploitability?
- Can we exploit properties of Leduc Poker to optimize parameters?
- Can we combine agents into an ensemble that minimizes exploitability further than its parts?

References

[1] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. Openspiel: A framework for reinforcement learning in games, 2019.

[2] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1729–1736. Curran Associates, Inc., 2008. URL http://papers.nips.cc/paper/3306-regret-minimization-in-games-with-incomplete-information.

Appendix

3.1 Time spent