Training Data Poisoning for Imperfect Information Games

Guy Aridor[1], Jisha Jacob[2], Natania Wolansky[2], and Iddo Drori[2]

[1]Department of Economics, Columbia University, New York, New York 10027, USA
[2]Department of Computer Science, Columbia University, New York, New York 10027, USA

Abstract

Most of the recent breakthrough work in artificial intelligence has been in developing agents that can play difficult multi-agent games such as chess, shogi, Go, or poker at super-human levels. However, are these agents susceptible to defeat by strategic agents with little computational power but with the ability to bias training? In real world chess and poker tournaments, grandmasters and expert players often use strategies of deception in practice rounds and early rounds of the tournament to confuse their opponents about the strategies they employ later on. This work explores how simple strategies in the game of Leduc Hold'em can be used to beat a sophisticated poker AI, DeepStack. We first implement agents that exhibit the behavioral biases that have been empirically observed in individuals playing poker. We then play these sub-optimal agents against an unbiased trained DeepStack and show how significantly DeepStack outperforms these traditional strategy profiles. We then consider the ability of an opponent to bias the training phase such that DeepStack is optimized to play against a particular strategy profile as opposed to approximating a Nash Equilibrium. Finally, by allowing for this biasing, we show that DeepStack can be defeated by a subset of strategy profiles if the player can change their strategy post-training. While DeepStack achieves nearly super-human performance, we conclude that DeepStack is susceptible to training poisoning.