Review-35

Fictitious Self-Play in Extensive-Form Games

Utilization of fFictitious Play (FP) as a learning mechanism has been essential. From the perspective of imperfect information games, fictitious learning allows the agent to leverage self and opponent's best responses towards optimal control. The work presents two variants of FP, namely full-width extensive-form fictitious play (XFP) and fictitious self-play (FSP). While XFP is realization-equivalent and inherits the convergence properties of FP, it is only applicable to behavioral strategies which restricts its scalability to higher dimensions. FSP addresses this limitation by replacing the fictitious operations of best response computation and average strategy updating with machine learning algorithms. FSP learns approximate best responses using Reinforcement Learning (RL) and strategies are averaged as a supervised learning task. Experiments conducted on imperfect-information poker games demonsrtate the suitable convergence of proposed approaches towards approximate Nash Equilibria.

The two propsed variants aim to enhance learning of agent strategies in order to improve convergence towards approximate Nash Equilibria. XFP demonstrates realization equivalent behavioral strategies which are equivalent to normal-form fictitious play. At each iteration, XFP computes a best response profile to the current strategies. It then uses the best response profile to update the average strategy profile. Updates require computations to be performed at all states irrespective of their relevance to the game. This results in the need for an extensive computational budget and leads to the formulation of FSP. Contrary to XFP, FSP is a machine-learning framework which implements generalized weakened FP in a sample-based fashion. The agent is provided with two types of memories, one stores opponents' behaviors and the other stores agent's own behaviors. The agent computes an approximate best response using RL off-policy from its memory of its opponent's behavior. Upon computation of best-response, the agent updates its own strategy using supervised learning from the memory of its own behavior. FSP mixes between the best response and average strategies using a mixing parameter for data generation.

XFP and FSP demonstrate suitable convergence towards approximate Nash Equilibria on Texas Hold'em Poker. FSP is further found to be scalable on the 60 card deck variant wherein XFP learns slowly as a result of the large computational budget requirement. While the variants demonsrtate theoretical suitability, they do not significantly highlight the improvements to FP. The work does not compare XFP and FSP to the baseline FP which would aid in understanding the extend of improvements provided by the methods. Furthermore, the work does not throw light on the overall performance of XFP and FSP in yielding higher payoffs. It would be interesting to compare the average performance of these methods to baseline FQI upon convergence.

XFP and FSP provide two novel directions for future work through the lens of imperfect information games. Firstly, the methods may be extended to large-scale games with varying amount of visible information to its players. Secondly, the trade-off between sound convergence and higher payoffs can be addressed as a result of the increasing complexity of control problem.

FP provides a suitable framework for learning in imperfect-information games. The work presents XFP and FSP as two variants of FP for approximating Nash Equilibria. XFP extends fictitious best response computation and average strategy updates to the extensive-form setting. The end result of XFP are realization-equivalent behavioral strategies at the cost of linearly scaling compute. FSP overcomes the scaling limitations of XFP by approximating best response using RL and average strategy updates using supervised learning. The proposed algorithms demonstate improved

convergence towards Nash Equilibria with the performance of FSP scalable to a larger deck size.