Review-39

Learning to Play No-Press Diplomacy with Best Response Policy Iteration

Advances in Reinforcement Learning (RL) have seen ample growth in games which do not necessarily represent complex many-agent interactions consisting of cooperation and competition. To address this challengs, the work considers the game of game of Diplomacy consisting of social dilemmas in many-agent scenarios. The work proposes the Sample Best Response (SBR) in conjunction with approximate policy iteration which closely resembles fictious play. Agents trained using these methods demonstrate improved performance and approximation to stable equilibria in comparison to state-of-the-art methods.

Various multi-agent methods focus on adversarial settings of competition. However, most practical scenarios are a complex mixture of cooperation and competition between multiple agents. Through this lens, the work aims to tackle Diplomacy, a 7-player board game which presents complex social dilemmas through simulatenous moves and large combinatorial action space. To solve the simpler version of No-Press Diplomacy, the paper presents the scalable SBR operator. SBR is a tractable approximation which leads to single-turn improvements. At each operation of SBR, the agent samples a small set of actions as its best response from a candidate policy. The candidate policy is evaluated against Monte-Carlo estimates of opponent's actions. As a result of this sampled relationship, strength of SBR directly depends on the strength of candidate's policy. SBR is applied to the latest policy and value which yields an improved policy. The improved policy is utilized to sample experiences of self-play which are used to fit a new policy and value. This iterative process of sampling and improving the current policy is referred to as Iterative Best Response (IBR) and proposes the Best Response Policy Iteration (BRPI).

SBR and BRPI are applied to the No-Press Diplomacy setting by making use of the improved DipNet architecture. The proposed methods when combined together demonstrate improved performance in comparison to the baseline methods on head-to-head and population-based comparison. Furthermore, the setup introduces meta-games, wherein the agent elects an 'AI Champion' to play on its behalf and then tries to match its score. Progression of IBR on meta-games demonstrates improvement in the quality of strategies and convergence towards stable equilibria solutions. The work further assesses epxloitability which presents 48% winrate. While the work presents a suitable large-scale application of RL to No-Press Diplomacy, it does not throw light on two aspects of its experimental setup. Firstly, the work does not highlight suitable transfer of policies among agents. Can the learned policies be transferred to other opponent agents midway in order to boost collaboration? Secondly, the work does not explain how the framework may be used to collaborate with other complex agents such as humans. Can the learned policies be paired-up with human players yo yield optimal results? The experimental setup could be improved to answer this question.

The setting of SBR and BPI presents several new directions for future work as identified in the paper. For instance, the work maybe extended towards (1) human-level performance in Diplomacy, (2) reduction in epxloitability of agents, (3) resoning about incentives, (4) learning to communicate intentions, (5) handling other variants of game and (5) gameplays against human players.