Enhancing Policy Gradient Algorithms With Search in Imperfect Information Games

Ondřej Kubíček
Czech Technical University in Prague, FEE, Dept. of CS, Artificial Intelligence Center
kubicon3@fel.cvut.cz

Abstract

Sequential decision-making under uncertainty in multi-agent environments is a fundamental problem in artificial intelligence. Games serve as a base model for these problems. Finding optimal plans in games that model real-world scenarios necessitates scalable algorithms. In games with perfect information, algorithms that use a combination of search and deep reinforcement learning can scale to arbitrary-sized games and achieve superhuman performance. In games with imperfect information, the situation is more challenging due to the nature of the search. This work aims to develop algorithms that use search but can scale into larger games than currently possible.

1 Introduction and Related Work

Most search algorithms for larger imperfect information games, like Poker, are built upon the Counterfactual Regret Minimization (CFR) with decomposition [Moravčík *et al.*, 2017; Schmid *et al.*, 2023]. CFR does not perform search in a traditional sense, but it iteratively approximates Nash equilibrium in two-player zero-sum games by minimizing regret of strategies. Even with this distinction, it is widely agreed that the CFR should be considered as a search algorithm. With decomposition, the search is performed only for a limited search depth in each encountered decision node. Notable algorithms that use this approach are Deepstack and Student of Games [Moravčík *et al.*, 2017; Schmid *et al.*, 2023]. These algorithms perform search both during train and test time in all encountered parts of the game. Search in training is necessary because the algorithm requires a complicated value function that predicts value based on the strategy employed in previous parts of the game. Any theoretically sound search in imperfect information games must simultaneously consider the whole set of game states that any player might consider to be the current state of the game. Games like Stratego or Dark Chess contain some parts with more than 10^{20}, which makes these algorithms not applicable. Also, sound search requires keeping multiple statistics throughout the game from the searches conducted in all previous parts.

A different approach is to use some form of abstraction, which makes the game smaller, and then it is possible to apply search algorithm [Čermák *et al.*, 2020]. Libratus used Poker-specific abstractions to reach a superhuman performance [Brown and Sandholm, 2018]. The Libratus used different abstractions in different parts of the game so that the search could use as many resources as possible. The problem with these abstractions is their domain-specificity. Outside of Poker, there have not been many successful applications of abstractions in large imperfect information games.

There is also a family of model-free algorithms, like RNaD, DREAM, or ARMAC, that are able to approximate Nash equilibrium without any search [Perolat *et al.*, 2022]. This makes them applicable to arbitrary-sized games. However, these algorithms fully rely on a trained policy network. It was shown that even in AlphaZero for perfect information games, using just the policy network leads to highly exploitable strategies [Wang *et al.*, 2022] and that the additional search improves these strategies to be less vulnerable to adversarial opponents. Similarly, the authors of RNaD had to incorporate multiple post-processing heuristics to achieve competitive performance in Stratego.

2 Multi-Agent Search With Policy Transformations

We have introduced an algorithm called Multi-agent Search with Policy Transformations (SePoT) [Kubíček *et al.*, 2024], which enhances trajectory-based policy-gradient algorithms like RNaD [Perolat *et al.*, 2022] with additional search during test time. The main idea of SePoT is to train an additional state critic network alongside the policy network, which, for each player, predicts the expected value of playing the policy from the policy network against some opponent's strategy. Such critic can then be used as a value function in a CFR search with an additional computational step [Brown *et al.*, 2018]. Training of the critic does not require any additional sampled trajectories besides those used for the policy training if the critic is trained with some off-policy predictor, like V-trace. This additional critic, alongside the policy network, provides sufficient information to perform safe search in any part of the game without the need to keep statistics from previous searches. This allows SePoT to be used even in games where the search cannot be performed in certain parts.

In order for critic to work as a good approximation of the value function, it is necessary to have a good coverage of opponents strategies. In the context of SePoT, this means that to any policy player can play, the oppponent can perform a best response. Since a best response can always be a pure strategy, having all pure strategies will cover this space perfectly [Brown *et al.*, 2018]. The amount of pure strategies in sequential games grows exponentially with the number of decisions. However, using only a handful of strategies still shows prominent results. Instead of learning all the opponent strategies, the SePoT learns only a single policy with the policy network, and simultaneously, it learns transformations of this policy for both players. These transformations shift the policy in a prominent direction, which are encountered during training.

3 Future Work

We have shown that SePoT is able to outperform the underlying policy-gradient algorithm and is scalable to larger games than algorithms like Deepstack, Libratus, or Student of Games, but some limitations still remain. We will now discuss these limitations.

Right now, the theoretical guarantees of the SePoTs search assume that the distance from the optimal value function is bounded, which is similar to the Student of Games. However, we have chosen the opponent strategies to be transformations of the policy based on the directions the policy network shifted throughout the training. This approach was chosen empirically, so the quality of the resulting strategies is not guaranteed. A more principled way to choose the opponent strategies that would contain some theoretical guarantees on the coverage of the strategy space would allow us to state proper theoretical bounds on the performance of SePoT.

The Student of Games always starts the search from the states present in the previous search. So, it always has access to all the possible current states. On the other hand, SePoT can start search in the middle of the game without using search anywhere previously, which was not possible with other algorithms. In such a scenario, starting a search from all possible current states is necessary. Reconstructing all these states is easy in some games like Stratego, Battleships, or Goofspiel, but in some other games like Dark Chess, it is computationally expensive. Furthermore, this is the only part of the algorithm that requires human-coded knowledge because the effective reconstruction differs between different games. In the future, we would like to tackle this issue using the diffusion model, which will generate all the states based on the given common knowledge information.

SePoT can follow the policy network in parts of the game where it is impossible to conduct search. However, from the experimental results, it is clear that the search is improving the policy. In games like Stratego, most parts of the game are intractable for search. Libratus did not use depth-limited solving, but it used abstractions to condense the game of Poker for search. One of the future directions is to define the quality of abstractions and then find a principled way to learn these good abstractions. Such learned abstracted game may have different dynamics than the original game. Extending model-learning ideas from MuZero [Schrittwieser *et al.*, 2020] to imperfect information games and immediately learning the abstracted game model could be a promising direction. The diffusion models for generating states from previous point can naturally be extended to predict these abstracted game states instead of the real ones.

Combining all the mentioned ideas would lead to a completely model-free algorithm, which can improve policy based on available resources during gameplay while still being scalable to arbitrary-sized two-player zero-sum games.

Acknowledgments Thanks to my supervisor doc. Mgr. Viliam Lisý, MSc., Ph.D. for his guidance and support on this project. This research is supported by Czech Science Foundation (GA22-26655S) and Grant Agency of the CTU in Prague (SGS22/168/OHK3/3T/13). Computational resources were supplied by (e-INFRA CZ LM2018140) supported by the MEYS of the Czech Republic and also (CZ.02.1.01/0.0/0.0/16 019/0000765).

References

[Brown and Sandholm, 2018] Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

[Brown *et al.*, 2018] Noam Brown, Tuomas Sandholm, and Brandon Amos. Depth-limited solving for imperfect-information games. *NeurIPS*, 31, 2018.

[Kubíček *et al.*, 2024] Ondřej Kubíček, Neil Burch, and Viliam Lisý. Look-ahead search on top of policy networks in imperfect information games, 2024.

[Moravčík *et al.*, 2017] Matej Moravčík, Martin Schmid, Neil Burch, et al. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[Perolat *et al.*, 2022] Julien Perolat, Bart De Vylder, Daniel Hennes, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

[Schmid *et al.*, 2023] Martin Schmid, Matej Moravčík, Neil Burch, et al. Student of games: A unified learning algorithm for both perfect and imperfect information games. *Science Advances*, 9(46):eadg3256, 2023.

[Schrittwieser *et al.*, 2020] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[Wang *et al.*, 2022] Tony Tong Wang, Adam Gleave, Nora Belrose, et al. Adversarial policies beat professional-level go ais. In *Deep Reinforcement Learning Workshop NeurIPS 2022*, 2022.

[Čermák *et al.*, 2020] Jiří Čermák, Viliam Lisý, and Branislav Bošanský. Automated construction of bounded-loss imperfect-recall abstractions in extensive-form games. *Artificial Intelligence*, 282:103248, 2020.