Continual Depth-limited Responses for Computing Counter-strategies in Sequential Games

Extended Abstract

David Milec
AI Center, FEE, CTU in Prague
Czech Republic
milecdav@fel.cvut.cz

Ondřej Kubíček
AI Center, FEE, CTU in Prague
Czech Republic
kubicon3@fel.cvut.cz

Viliam Lisý
AI Center, FEE, CTU in Prague
Czech Republic
viliam.lisy@agents.fel.cvut.cz

ABSTRACT

In zero-sum games, the optimal strategy is well-defined by the Nash equilibrium. However, it is overly conservative when playing against suboptimal opponents and it can not exploit their weaknesses. Limited look-ahead game solving in imperfect-information games allows superhuman play in massive real-world games such as Poker, Liar's Dice, and Scotland Yard. However, since they approximate Nash equilibrium, they tend to only win slightly against weak opponents. We propose theoretically sound methods combining limited look-ahead solving with an opponent model, in order to 1) approximate a best response in large games or 2) compute a robust response with control over the robustness of the response. Both methods can compute the response in real time to previously unseen strategies. We present theoretical guarantees of our methods. We show that existing robust response methods do not work combined with limited look-ahead solving of the shelf, and we propose a novel solution for the issue. Our algorithm performs significantly better than multiple baselines in smaller games and outperforms state-of-the-art methods against SlumBot.

KEYWORDS

large games; approximating best response; robust response; opponent exploitation; imperfect information; depth limited solving

ACM Reference Format:
David Milec, Ondřej Kubíček, and Viliam Lisý. 2024. Continual Depth-limited Responses for Computing Counter-strategies in Sequential Games: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION [14]

Adapting to suboptimal opponents can significantly improve our winnings, even in zero-sum games. There is substantial work on modeling and adapting to opponents both in theory [2, 6, 8, 12, 13, 16, 18] and also applied in real-world [1, 4].

However, the current methods are limited by the size of the game and need the game to fit in the memory. Decomposition is a key method to solve games too large to fit in the memory [3]. An algorithm successfully applying decomposition in imperfect-information games is the continual resolving [15]. Here, we extend it to exploit opponents effectively.

We create algorithms to both fully exploit known opponents and exploit known opponents while staying robust to worst-case adversaries. We call the methods continual depth-limited best response (CDBR) and continual depth-limited restricted Nash response (CDRNR). We prove the theoretical properties of the algorithms and show that CDBR exploits SlumBot significantly more than the previous state-of-the-art. Furthermore, CDRNR can exploit opponents more than previous state-of-the-art and has stronger properties in balancing robustness and exploitation.

2 FULLY EXPLOITING THE OPPONENT

CDBR is an algorithm that focuses on fully exploiting the opponent. We need a value function since it is based on continual resolving [9, 15]. This value function summarizes what will happen in the future parts of the game. To compute the exact best response, we would need to capture the opponent's strategy in the value function, resulting in a need for a separate value function for any possible opponent. We use a single optimal value function to avoid this, assuming both players continue optimally. We fix the opponent in the look-ahead part of the game we have in the memory and compute the strategy against it using the optimal value function. We show both theoretical results showing we will get at least the value of the game and also strong empirical performance.

3 SAFE MODEL EXPLOITATION

While CDBR maximizes the exploitation of the fixed opponent model, it allows the player to be exploited if the opponent deviates from the assumed strategy.

Combination of CDBR and Nash Equilibrium

The combination of CDBR and Nash equilibrium (CDBR-NE) is the first approach to limit exploitability. We can simultaneously compute both strategies using depth-limited solving and do a linear combination in every decision node. The gain and exploitability of the resulting linear combination is then a linear combination of the gain and exploitability of the combined strategies. We can achieve the desired exploitability or gain by tuning the parameter of the linear combination, and the algorithm is only two times slower than the CDBR since we need to find the Nash equilibrium separately and perform CDBR. The required value function is the same for both parts and is still the same as in standard continual resolving.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Continual Depth-limited RNR

CDBR-NE is safe, but [7] shows we can get a much better trade-off between gain and exploitability using RNR as it recovers the optimal Pareto set of ϵ-safe best responses [11]. RNR gives us better safety control by linking the allowed exploitability to the achieved gain. We combine depth-limited solving with RNR to create CDRNR.

Description of Restricted Nash Response. For CDRNR, we first need to explain the RNR method briefly [7]. RNR begins with an initial chance action with two outcomes, where the likelihood of the outcomes determines the robustness parameter. We copy the whole game after both of the outcomes. The player computing the RNR can not observe which outcome was chosen. In one part of the game, the opponent is fixed; in the other, it behaves rationally.

Continual Depth-limited Restricted Nash Response. The first problem of CDRNR is the value function. At the depth-limit, we have a differently sized vector for the value function. However, the structure of the value function can stay the same as in the continual resolving. When we apply it, we need to combine the opponent's strategies from the fixed and rational responses and re-weight the results. The second problem is the robustness. Continual resolving uses a gadget, an addition to the top of the game [3]. This modification ensures the robustness of the strategy. Our fundamental result is that previously used gadgets all fail with suboptimal opponents, and the minimal working gadget is what we call a full gadget. The full gadget keeps all the histories from the root to the current part of the game and uses the value function to evaluate the branches to other parts. We show that we always achieve at least value of the game against the chosen opponent. We also show a bound on the exploitability, which we connect directly to what we obtain from the chosen opponent. We also beat the previous method and are close to the RNR, the perfect response for a given parameter.

Exploitability of Robust Responses

We report both gain and exploitability for CDRNR on Leduc Hold'em. Results in Figure 1 show that the proven bound on exploitability works in practice and is very loose. For example, with $p = 0.5$, the bound on the exploitability is the gain itself, but the algorithm rarely reaches even a tenth of the gain in exploitability. Hence, the CDRNR, similarly to RNR, can significantly exploit the opponent without significantly raising its exploitability with a well set p.

We compare against the best possible Nash equilibrium computed by a linear program. It is the theoretical limit of maximal gain, which does not allow exploitability. We can see we can gain over twice as much, with exploitability still almost zero. Safe exploitation search (SES) is a previous method that only uses fixed opponent reaches. It also uses the standard gadget, which can not give strong guarantees [10].

Playing SlumBot

We tested our method in HUNL against SlumBot [5], which is a publicly available abstraction-based bot commonly used for benchmarking. We used a fold, call, pot, and all-in (FCPA) abstraction for CDBR. CDBR significantly outperforms ABR and LBR, and we report the results in Table 1. Authors in [17] also use FCPA for their method but did not report a confidence interval for the results.

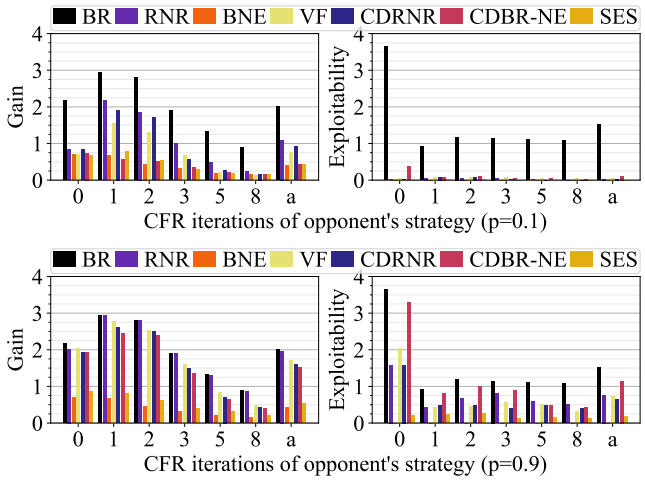

Figure 1: Gain and exploitability of BR, RNR, best Nash equilibrium (BNE), CDBR-NE, SES, and CDRNR in Leduc Hold'em against CFR strategies with a small number of iterations with different p values. The a represents the average of the other values. VF is CDRNR using an imperfect value function.

	ABR	LBR	CDBR
Win-rate [mbb/h]	1259 ± ?	1388 ± 150	1774 ± 137

Table 1: Comparison of CDBR with LBR and ABR against SlumBot. Results are reported in milibigblinds per hand (mbb/h) with 95% confidence intervals. (Authors of ABR did not report a confidence interval.)

4 CONCLUSION

We propose new algorithms to compute responses in large imperfect-information games, creating the best performing theoretically sound robust response applicable to games that require decomposition. We empirically evaluate the algorithms on multiple games. We show that CDBR outperforms LBR in both Leduc and HUNL, and we show that CDBR performs significantly better against SlumBot than any other previous method. Finally, we show that CDRNR outperforms SES in any game and can achieve over half the possible gain without almost any exploitability.

ACKNOWLEDGEMENTS

This research was supported by Czech Science Foundation (grant no. GA22-26655S) and the Grant Agency of the Czech Technical University in Prague, grant No. SGS22/168/OHK3/3T/13. Computational resources were supplied by "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic and also the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics". We also greatly appreciate the help of Eric Jackson, who provided the SlumBot strategy and helped with the experiments.

REFERENCES

[1] Bo An, Fernando Ordóñez, Milind Tambe, Eric Shieh, Rong Yang, Craig Baldwin, Joseph DiRenzo III, Kathryn Moretti, Ben Maule, and Garrett Meyer. 2013. A deployed quantal response-based patrol planning system for the US Coast Guard. *Interfaces* 43, 5 (2013), 400–420.

[2] Nolan Bard, Michael Johanson, Neil Burch, and Michael Bowling. 2013. Online implicit agent modelling. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. 255–262.

[3] Neil Burch, Michael Johanson, and Michael Bowling. 2014. Solving imperfect information games using decomposition. In *Twenty-eighth AAAI conference on artificial intelligence*.

[4] Fei Fang, Thanh Hong Nguyen, Rob Pickles, Wai Y Lam, Gopalasamy R Clements, Bo An, Amandeep Singh, Brian C Schwedock, Milind Tambe, and Andrew Lemieux. 2017. PAWS-A Deployed Game-Theoretic Application to Combat Poaching. *AI Magazine* 38, 1 (2017), 23–36.

[5] Eric Griffin Jackson. 2017. Targeted cfr. In *Workshops at the thirty-first AAAI conference on artificial intelligence*.

[6] Michael Johanson and Michael Bowling. 2009. Data biased robust counter strategies. In *Artificial Intelligence and Statistics*. 264–271.

[7] Michael Johanson, Martin Zinkevich, and Michael Bowling. 2008. Computing robust counter-strategies. In *Advances in neural information processing systems*. 721–728.

[8] Kevin B Korb, Ann Nicholson, and Nathalie Jitnah. 2013. Bayesian poker. *arXiv preprint arXiv:1301.6711* (2013).

[9] Vojtěch Kovařík, Dominik Seitz, Viliam Lisý, Jan Rudolf, Shuo Sun, and Karel Ha. 2023. Value functions for depth-limited solving in zero-sum imperfect-information games. *Artificial Intelligence* 314 (2023), 103805.

[10] Mingyang Liu, Chengjie Wu, Qihan Liu, Yansen Jing, Jun Yang, Pingzhong Tang, and Chongjie Zhang. 2022. Safe Opponent-Exploitation Subgame Refinement. In *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=YpHb0IVJu92

[11] Peter McCracken and Michael Bowling. 2004. Safe strategies for agent modelling in games. In *AAAI Fall Symposium on Artificial Multi-agent Learning*. 103–110.

[12] Richard Mealing and Jonathan L Shapiro. 2015. Opponent modeling by expectation–maximization and sequence prediction in simplified poker. *IEEE Transactions on Computational Intelligence and AI in Games* 9, 1 (2015), 11–24.

[13] David Milec, Jakub Černý, Viliam Lisý, and Bo An. 2021. Complexity and Algorithms for Exploiting Quantal Opponents in Large Two-Player Games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 5575–5583.

[14] David Milec, Ondřej Kubíček, and Viliam Lisý. 2021. Continual Depth-limited Responses for Computing Counter-strategies in Sequential Games. *arXiv preprint arXiv:2112.12594* (2021).

[15] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. 2017. Deepstack: Expert-level artificial intelligence in Heads-Up No-Limit Poker. *Science* 356, 6337 (2017), 508–513.

[16] Finnegan Southey, Michael P Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. 2012. Bayes' bluff: Opponent modelling in poker. *arXiv preprint arXiv:1207.1411* (2012).

[17] Finbarr Timbers, Edward Lockhart, Marc Lanctot, Martin Schmid, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. 2020. Approximate exploitability: Learning a best response in large games. *arXiv preprint arXiv:2004.09677* (2020).

[18] Zhe Wu, Kai Li, Enmin Zhao, Hang Xu, Meng Zhang, Haobo Fu, Bo An, and Junliang Xing. 2021. L2E: Learning to Exploit Your Opponent. *arXiv preprint arXiv:2102.09381* (2021).