Bridging the Gap between Partially Observable Stochastic Games and Sparse POMDP Methods

Tyler Becker
University of Colorado Boulder
Boulder, United States
tyler.becker-1@colorado.edu

Zachary Sunberg
University of Colorado Boulder
Boulder, United States
zachary.sunberg@colorado.edu

ABSTRACT

Many real-world decision problems involve the interaction of multiple self-interested agents with limited sensing ability. The partially observable stochastic game (POSG) provides a mathematical framework for modeling these problems, however solving a POSG requires difficult reasoning over two critical factors: (1) information revealed by partial observations and (2) decisions other agents make. In the single agent case, partially observable Markov decision process (POMDP) planning can efficiently address partial observability with particle filtering. In the multi-agent case, extensive form game solution methods account for other agent's decisions, but preclude state-belief approximation. We propose a unifying framework that combines POMDP-inspired state distribution approximation and game-theoretic equilibrium search on information sets. This paper lays a theoretical foundation for the approach by bounding errors due to belief approximation, and empirically demonstrates effectiveness with a numerical example. The new approach enables planning in POSGs with very large state spaces, paving the way for reliable autonomous interaction in real-world physical environments and complementing multi-agent reinforcement learning.

KEYWORDS

Game Theory, Imperfect Information, Particle Filter, Counterfactual Regret, POSG, POMG, POMDP

ACM Reference Format:
Tyler Becker and Zachary Sunberg. 2025. Bridging the Gap between Partially Observable Stochastic Games and Sparse POMDP Methods. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

This paper addresses game-theoretic planning in large, partially observable state spaces, where both state uncertainty and interaction uncertainty drive complex behaviors. Agents may cooperate, compete adversarially, or have general-sum objectives, requiring strategic reasoning beyond standard POMDP frameworks.

Existing methods either focus on state uncertainty (POMDP solvers) or interaction uncertainty (game-theoretic approaches), but handling both remains a challenge. We introduce the Conditional Distribution Information Set Tree (CDIT), a structure that enables belief approximation and multi-agent reasoning in extensive-form games. By representing a POSG in a way that existing game-solving algorithms such as CFR can operate on, CDIT makes it possible to find equilibria in continuous-state settings. We show theoretically that a Nash equilibrium found using this approximation remains close to the true equilibrium, without any direct dependence on state space size.

Finally, we demonstrate CDIT's effectiveness in a continuous-state tag game, where existing POMDP and extensive-form game methods fail.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 BACKGROUND

A partially observable stochastic game (POSG), also called a partially observable Markov game (POMG), models multi-agent decision-making under uncertainty, where players maximize individual utilities based on partial observations [1, 9]. Unlike single-agent optimization problems, POSGs lack globally optimal solutions; instead, equilibria such as Nash equilibria serve as solution concepts.

Imperfect information extensive-form games (EFGs) offer an alternative framework for multi-agent decision-making but differ structurally from POSGs in their sequential nature, terminal rewards, and information set representation of uncertainty. While zero-sum EFGs can be solved via regret minimization techniques such as counterfactual regret minimization (CFR) [20], these methods struggle with physical-world domains where belief approximation is necessary.

Tree search and belief-state planning techniques, common in single-agent POMDPs [7, 13, 16, 18], can extend to cooperative multi-agent settings (decentralized POMDPs) [19], but fail in general-sum games where mixed strategies may be required. POSG beliefs depend on other agents' policies, making equilibrium computation challenging without exact belief updates, which are impractical in large or continuous spaces [5].

Existing game-theoretic solvers either leverage deep learning [6, 8, 12, 17]—which lacks theoretical guarantees—or rely on explicit reach probabilities [3, 4, 11, 15], limiting their applicability to structured problems. Our approach circumvents these issues by efficiently approximating expected utilities while preserving theoretical guarantees, making it scalable to large, continuous domains without reliance on extensive hyperparameter tuning.

3 CONDITIONAL DISTRIBUTION INFORMATION SET TREES

In order to overcome the limitations above, we define a new tree structure called the conditional distribution information set tree (CDIT) that combines history-conditioned state distributions, similar to POMDP beliefs, with information sets similar to those in EFGs.

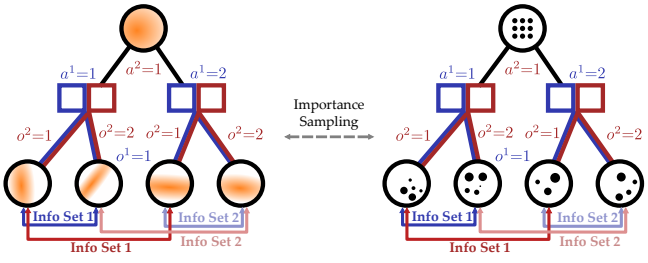

Figure 1: Illustration of a CDIT (left) and its particle approximation (right) for a POSG with $\mathcal{A}^1$=O^2={1,2}, $\mathcal{A}^2$=O^1={1}.

The base structure of a CDIT is a joint conditional distribution tree consisting of alternating layers of joint action nodes (rectangles in Fig. 1) and joint observation nodes (unfilled circles in Fig. 1). The history for a node is the sequence of joint actions and joint observations on the path to that node from the root. Each depth d observation node has an associated state distribution, b_d, conditioned on the history up to that point. This distribution can be calculated exactly using Bayes' rule. However, CDITs are most scalable when this belief is approximated. Any approximation, for example an extended Kalman filter or Gaussian mixture model can be used, but this work focuses on a *particle CDITs*, where each distribution is represented by C particles (filled circles in Fig. 1). Particles are propagated by sampling the joint transition distribution, and particle weights are updated according to observation probability.

Since the distributions in the tree described above are conditioned on joint histories, they contain more information than any one player has at a given depth. In order to limit the information that policies can be conditioned on, distribution nodes corresponding to histories that are indistinguishable to a player are grouped together into information sets for each player. Specifically, two nodes are in the same information set for a player if all actions and observations for that player in the history leading up to that node are identical. This grouping is similar to the information set concept in EFGs. However, while EFGs may arbitrarily group states into information sets, CDITs by definition group nodes according to the criterion above. The combination of a joint conditional distribution tree and history-based information sets constitutes a CDIT. An information set in a CDIT can be interpreted as a summary of the information about the state implied by the history observed by an agent, without assuming anything about other agents' policies. We provide guarantees for using external sampling counterfactual regret minimization (ESCFR) to efficiently traverse the CDIT and synthesize a policy.

4 CONVERGENCE GUARANTEES FOR APPROXIMATE NASH EQUILIBRIA ON CDITS

When using a particle CDIT to approximate a game, a crucial question is whether equilibria computed on the CDIT, for example with the ESCFR algorithm, converge to equilibria in the original game as the number of particles increases.

By leveraging recent work in POMDP particle approximation [10], our guarantees come in the form of a concentration bound that is parametric in the solving method. It demonstrates that a solver with guarantees only for smaller finite games now may enjoy guarantees for large or even continuous state spaces when using the CDIT.

We separate the convergence guarantees into three parts [2]. First, we show that the suboptimality of a solution calculated using an approximate game is bounded when applied to the true game. Then, we bound utility approximation error of this approximate game. Next, we show that using a sampled subset of the strategies and observations in the approximate game is sufficient to solve the approximate game. Finally, we bound the suboptimality of a sparse ESCFR solution.

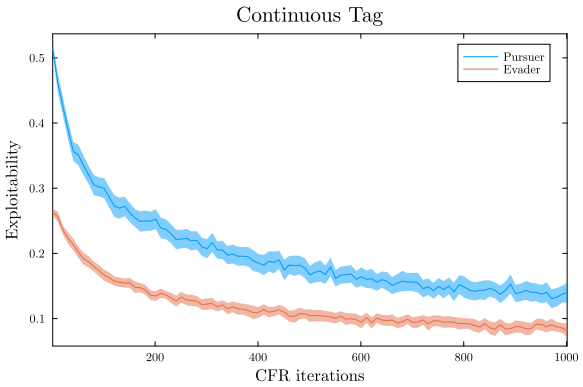

Figure 2: Continuous Tag exploitability; 3σ standard error bounds shaded

5 NUMERICAL EXPERIMENTS

To demonstrate the effectiveness of our solver, we construct a game of partially observable tag in 2-dimentional continuous state space for each agent (4 dimensions total) with discrete finite actions and observations.

To elucidate the adversarial nature of the interaction uncertainty in these zero-sum games, we quantify suboptimality via exploitability, which we define to be how much utility an opponent is able to take away, should they know the produced policy exactly. This exploitability over the course of solving is given in Fig. 2.

6 CONCLUSION

This paper introduces a novel approach to solving POSGs by integrating imperfect information game methods with POMDP-based distribution approximations. This enables low-exploitability solutions in continuous state spaces and large observation spaces. While our method improves scalability by reducing search complexity, it remains intractable for large action spaces. Future work could explore model-free deep reinforcement learning to further sparsify CDIT sampling and generalize results. Additionally, while our approach lacks online replanning, extensions based on [14, 15] could enable adaptive strategy refinement.

ACKNOWLEDGMENTS

This work is supported by the Air Force Office of Scientific Research (AFOSR), Grant number FA9550-23-1-0726.

REFERENCES

[1] Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. 2023. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press. https://www.marl-book.com

[2] Tyler Becker and Zachary Sunberg. 2024. Bridging the Gap between Partially Observable Stochastic Games and Sparse POMDP Methods. *arXiv preprint arXiv:2405.18703* (2024).

[3] Noam Brown and Tuomas Sandholm. 2017. Safe and nested subgame solving for imperfect-information games. *Advances in neural information processing systems* 30 (2017).

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

[5] Aurélien Delage, Olivier Buffet, Jilles S Dibangoye, and Abdallah Saffidine. 2023. HSVI can solve zero-sum partially observable stochastic games. *Dynamic Games and Applications* (2023), 1–55.

[6] Arnaud Fickinger, Hengyuan Hu, Brandon Amos, Stuart Russell, and Noam Brown. 2021. Scalable online planning via reinforcement learning fine-tuning. *Advances in Neural Information Processing Systems* 34 (2021), 16951–16963.

[7] Neha P. Garg, David Hsu, and Wee Sun Lee. 2019. DESPOT-α: Online POMDP Planning With Large State And Observation Spaces. In *Robotics: Science and Systems*.

[8] Hengyuan Hu, Adam Lerer, Noam Brown, and Jakob Foerster. 2021. Learned belief search: Efficiently improving policies in partially observable settings. *arXiv preprint arXiv:2106.09086* (2021).

[9] Mykel J. Kochenderfer and Tim A. Wheeler. 2019. *Algorithms for Optimization*. MIT Press.

[10] Idan Lev-Yehudi, Moran Barenboim, and Vadim Indelman. 2024. Simplifying Complex Observation Models in Continuous POMDP Planning with Probabilistic Guarantees and Practice. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 20176–20184.

[11] Weiming Liu, Haobo Fu, Qiang Fu, and Yang Wei. 2023. Opponent-limited online search for imperfect information games. In *International Conference on Machine Learning*. PMLR, 21567–21585.

[12] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, 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.

[13] David Silver and Joel Veness. 2010. Monte-Carlo Planning in Large POMDPs. In *Advances in Neural Information Processing Systems*. 2164–2172. http://papers.nips.cc/paper/4031-monte-carlo-planning-in-large-pomdps.pdf

[14] Samuel Sokota, Gabriele Farina, David J Wu, Hengyuan Hu, Kevin A Wang, J Zico Kolter, and Noam Brown. 2023. The Update-Equivalence Framework for Decision-Time Planning. *arXiv preprint arXiv:2304.13138* (2023).

[15] Christopher Solinas, Doug Rebstock, Nathan Sturtevant, and Michael Buro. 2024. History filtering in imperfect information games: algorithms and complexity. *Advances in Neural Information Processing Systems* 36 (2024).

[16] Zachary Sunberg and Mykel Kochenderfer. 2018. Online algorithms for POMDPs with continuous state, action, and observation spaces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 28. 259–263.

[17] Michal Šustr, Vojtech Kovarík, and Viliam Lisy. 2021. Particle value functions in imperfect information games. In *AAMAS Adaptive and Learning Agents Workshop*.

[18] Nan Ye, Adhiraj Somani, David Hsu, and Wee Sun Lee. 2017. DESPOT: Online POMDP Planning with Regularization. *Journal of Artificial Intelligence Research* 58 (2017), 231–266.

[19] Kaiqing Zhang, Erik Miehling, and Tamer Başar. 2019. Online Planning for Decentralized Stochastic Control with Partial History Sharing. In *American Control Conference (ACC)*. IEEE, 3544–3550.

[20] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2007. Regret minimization in games with incomplete information. *Advances in neural information processing systems* 20 (2007), 1729–1736.