MEX Algorithm in HAI

BY JACK YANSONG LI
University of Illinois Chicago
Email: yli340@uic.edu

1 Introduction

Training an AI agent capable of cooperating with various types of humans stands as a central challenge in human-AI Interaction (HAI). This problem proves to be difficult because different humans can create varied environments for the AI agent to navigate. Additionally, the AI agent cannot presume rational behavior from humans within the collaboration setting [16]. Many studies in the realm of multi-agent reinforcement learning have primarily focused on centralized settings [63][61][24], which becomes problematic in scenarios lacking a feasible central control for individual agents.

Consequently, collaborating with unknown humans without predetermined communication or coordination guidelines becomes important, giving rise to a research area named ad-hoc teamwork [41][6][52]. A pivotal subtask within the domain of ad-hoc teamwork is opponent modeling [5][4][11][59][30], a concept that, in the context of Human-AI (HAI) interactions, primarily entails modeling human behaviors and policies.

In this paper, we tackle a problem in the Human-AI (HAI) interaction domain where there are two participants: a human agent and an AI agent. The human agent keeps her policy hidden from the AI. However, the AI starts with a few initial guesses about the human's possible policies. The main goal of the AI is to figure out the actual policies the human agent is using. We categorize this approach as falling within several known frameworks such as latent Markov Decision Processes (MDP) [36][28][14] and multi-task Reinforcement Learning (RL) [38] [55] [14]. We will explore these and others in greater detail in the related work.

To model the HAI problem, we use an episodic Markov decision process where the transitions and rewards are influenced by the policy that the human agent keeps secret. The AI agent begins with a set of initial guesses about the human's policy, grouped together in a finite hypothesis set $\mathcal{H}$. We discuss loosening the limitation of this finite set in a later section (refer to the section on the infinite hypothesis set). It's assumed that the actual policy the human is using is a part of set $\mathcal{H}$, an idea referred to as the realization assumption. We'll talk about easing this assumption in another section (see the section on the realization assumption).

We used the Maximize to Explore (MEX) algorithm mentioned in [39] to tackle the episodic MDP with an undefined hypothesis set. We confirm a sub-linear regret outcome in Section (finite regret). Moreover, we found that the MEX algorithm can naturally decrease the size of the hypothesis set by grouping together human policies that are of the same type, allowing for a regret boundary that is smaller than the upper limit noted in [39]. The definition for policies be of the same type is introduced in Section (same type). Furthermore, we applied the MEX algorithm with an infinite hypothesis set that encompasses the true policy. We demonstrated that utilizing MEX with a finite hypothesis set, which contains a policy nearly identical to the true policy in the infinite hypothesis set, can still achieve sub-linear regret that tends to a value that close to an optimal value. This aspect is elaborated in Section (infinite hypothesis set).

In our experiment, we developed a simplified environment of the Overcooked-AI [16], where agents are required to engage in a series of actions such as cooking, waiting, and delivering food. The simplified version of Overcooked-AI, focusing exclusively on the food delivery task. This simplification is essential as it enables us to focus clearly on the main challenges posed by the original environment but significantly reduce the size of both state and action spaces, leading to a considerable decrease in computational complexity.

We created the finite hypothesis set using the best response dynamics method [54], where agents constantly modify their policies to best respond to the policies observed from other agents. In addition to this, numerous studies have explored various approaches to develop a finite hypothesis set $\mathcal{H}$ [46][54]. The infinite hypothesis set is defined as an open cover of the generated finite hypothesis set, which mimic the situation that human agents' bounded rationality.

We compared our algorithm with Q-learning with UCB exploration [32][19], Upper confidence bound [37], optimistic posterior sampling [65], and UCRL2 [8] algorithms. Our results shows that (added experiment)

2 Related Work

Previous research in the Human-AI (HAI) field tends to model human policy as a policy that closely aligns with the AI agent, with efforts to parameterize this closeness [42]. To evaluate these algorithms, several benchmark environments are available that aid in analyzing cooperative human-AI interaction tasks, including platforms like the two-player cooperative Atari game [57], bridge card game [40], and Overcooked-AI [16][54].

Human-AI Interaction: Previous research in the Human-AI (HAI) field model human policy as a policy that closely aligns with the AI agent, with efforts to parameterize this closeness [42]. Additionally, there are studies in Meta Reinforcement Learning (Meta RL) that work on deciphering the MDP the AI agent encounters, inherently learning the human agent's policy, since the structure of this MDP is influenced by the human agent's choices. To evaluate these algorithms, several benchmark environments are available that aid in analyzing cooperative human-AI interaction tasks, including platforms like the two-player cooperative Atari game [57], bridge card [40], and Overcooked-AI [16][54].

Human agent generation: Collecting human policies can be notably costly. Previous studies have developed methods for more efficient human policies generation. One such method utilizes an algorithm that identifies and selects policies based on a measure defined by the diversity of the best responses these policies can offer. The algorithm then maximizes this measure to find the policies that provide a diverse set of best responses [46]. Another strategy formulates human policies by running best response dynamics [54].

Ad-Hoc teamworks: Our work is closely related to ad-hoc teamworks [41][6][52] , especially the opponent modeling subtask. Barrett et. al. [11] introduces PLASTIC-Model and PLASTIC-Policy algorithms, the formal algorithm models the team-member by its transition dynamics and the latter models team-member by its policy. He et.al. [30] models the human agent's policy as a deep neural network.

YL: needs more time to read [5][4][59]

Partially observable Markov decision process (POMDP): The foundation of our problem is closely related to the partially observable Markov decision process [7][50], since each human policy in the hypothesis set can be viewed as a latent variable of the POMDP. The POMDP problem where we have latent variables are called latent MDPs (LMDP) [36]. LMDP has few different names, such as contextual decision process [28], multi-model MDP [51], multi-task RL [38][55][14], MOMDP/hidden model MDP [43][17][20][23], and concurrent MDP [15]. Beyond original POMDP, there are also some other settings that can cover our problem, such as interactive-POMDP [29][25], Augmented Bayes-Adaptive MDP (BAMDP) [58][22][26]. The model-based RL with UCB exploration algorithms [53][9] is also related to our setting.

MEX related algorithms: Our algorithm is based on MEX [39], in each episode, the algorithm chooses a human policy from the hypothesis set. On the other hand, the posterior sampling algorithms [56][49][65][1][64][2][3] updates a belief over the hypothesis set in each episode and draws a policy based on the current belief. Some methods like OLIVE [31] eliminates policy from current hypothesis set in each episode. Additionally, there is a method that trains one policy that is robust for all possible human policies in the hypothesis set [13].

MDP structure assumptions: Our regret analysis is based on the low generalized eluder coefficient assumption [65], which is a weaker assumption than low Eluder dimension/Bellman eluder dimension [44][34], low Bellman rank [31], Bellman completeness [62], Bilinear classes [21], and linear MDP structure [60][35]. The environment we used in the experiment is a tabular MDP which satisfies the low Bellman eluder dimension assumption [34]. Also, the regret analysis of infinite hypothesis set are related to agnostic online learning [12][48].

Bibliography

[1] Alekh Agarwal and Tong Zhang. Model-based RL with Optimistic Posterior Sampling: Structural Conditions and Sample Complexity. oct 2022.

[2] Alekh Agarwal and Tong Zhang. Non-Linear Reinforcement Learning in Large Action Spaces: Structural Conditions and Sample-efficiency of Posterior Sampling. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 2776-2814. PMLR, jun 2022.

[3] Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. mar 2020.

[4] Stefano V. Albrecht, Jacob W. Crandall, and Subramanian Ramamoorthy. Belief and Truth in Hypothesised Behaviours. *Artificial Intelligence*, 235:63-94, jun 2016.

[5] Stefano V. Albrecht and Subramanian Ramamoorthy. A Game-Theoretic Model and Best-Response Learning Method for Ad Hoc Coordination in Multiagent Systems. jun 2015.

[6] Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66-95, may 2018.

[7] K.J Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174-205, feb 1965.

[8] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[9] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. jul 2017.

[10] Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement Learning of POMDPs using Spectral Methods. may 2016.

[11] Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence*, 242:132-171, jan 2017.

[12] Shai Ben-David, D. Pál, and S. Shalev-Shwartz. Agnostic Online Learning. In *Annual Conference Computational Learning Theory*. 2009.

[13] Luca F. Bertuccelli, Albert Wu, and Jonathan P. How. Robust Adaptive Markov Decision Processes: Planning with Model Uncertainty. *IEEE Control Systems Magazine*, 32(5):96-109, oct 2012.

[14] Emma Brunskill and Lihong Li. Sample Complexity of Multi-task Reinforcement Learning. sep 2013.

[15] Peter Buchholz and Dimitri Scheftelowitsch. Computation of weighted sums of rewards for concurrent MDPs. *Math Meth Oper Res*, 89(1):1-42, feb 2019.

[16] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the Utility of Learning about Humans for Human-AI Coordination. Oct 2019.

[17] Iadine Chades, Josie Carwardine, Tara Martin, Samuel Nicol, Regis Sabbadin, and Olivier Buffet. MOMDPs: A Solution for Modelling Adaptive Management Problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):267-273, 2012.

[18] Rujikorn Charakorn, Poramate Manoonpong, and Nat Dilokthanakul. Learning to Cooperate with Unseen Agent via Meta-Reinforcement Learning. nov 2021.

[19] Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with UCB Exploration is Sample Efficient for Infinite-Horizon MDP. sep 2019.

[20] Finale Doshi-Velez and George Konidaris. Hidden Parameter Markov Decision Processes: A Semiparametric Regression Approach for Discovering Latent Task Parametrizations. aug 2013.

[21] Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear Classes: A Structural Framework for Provable Generalization in RL. jul 2021.

[22] Michael O'Gordon Duff. Optimal learning: Computational procedures for Bayes -adaptive Markov decision processes. *Doctoral Dissertations Available from Proquest*, pages 1-247, jan 2002.

[23] A. Fern, S. Natarajan, K. Judah, and P. Tadepalli. A Decision-Theoretic Model of Assistance. *Journal of Artificial Intelligence Research*, 50:71-104, may 2014.

[24] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual Multi-Agent Policy Gradients. dec 2017.

[25] P. J. Gmytrasiewicz and P. Doshi. A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research*, 24:49-79, jul 2005.

[26] Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1025-1033. Red Hook, NY, USA, dec 2012. Curran Associates Inc.

[27] Z. Guo, Shayan Doroudi, and E. Brunskill. A PAC RL Algorithm for Episodic POMDPs. In *International Conference on Artificial Intelligence and Statistics*. May 2016.

[28] Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov Decision Processes. feb 2015.

[29] Yanlin Han and Piotr Gmytrasiewicz. Learning Others' Intentional Models in Multi-Agent Settings Using Interactive POMDPs. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[30] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent Modeling in Deep Reinforcement Learning. sep 2016.

[31] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual Decision Processes with Low Bellman Rank are PAC-Learnable. dec 2016.

[32] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning Provably Efficient? jul 2018.

[33] Chi Jin, Sham M. Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-Efficient Reinforcement Learning of Undercomplete POMDPs. oct 2020.

[34] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. jul 2021.

[35] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 2137-2143. PMLR, jul 2020.

[36] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. RL for Latent MDPs: Regret Guarantees and a Lower Bound. feb 2021.

[37] T. L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4-22, mar 1985.

[38] Yao Liu, Zhaohan Guo, and Emma Brunskill. PAC Continuous State Online Multitask Reinforcement Learning with Identification. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, pages 438-446. Richland, SC, may 2016. International Foundation for Autonomous Agents and Multiagent Systems.

[39] Zhihan Liu, Miao Lu, Wei Xiong, Han Zhong, Hao Hu, Shenao Zhang, Sirui Zheng, Zhuoran Yang, and Zhaoran Wang. One Objective to Rule Them All: A Maximization Objective Fusing Estimation and Planning for Exploration. may 2023.

[40] Edward Lockhart, Neil Burch, Nolan Bard, Sebastian Borgeaud, Tom Eccles, Lucas Smaira, and Ray Smith. Human-Agent Cooperation in Bridge Bidding. nov 2020.

[41] Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. A Survey of Ad Hoc Teamwork Research. In Dorothea Baumeister and Jörg Rothe, editors, *Multi-Agent Systems*, volume 13442, pages 275-293. Springer International Publishing, Cham, 2022.

[42] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36(5-7):618-634, jun 2017.

[43] Sylvie C. W. Ong, Shao Wei Png, David Hsu, and Wee Sun Lee. Planning under Uncertainty for Robotic Tasks with Mixed Observability. *Int. J. Rob. Res.*, 29(8):1053-1068, jul 2010.

[44] Ian Osband and Benjamin Van Roy. Model-based Reinforcement Learning and the Eluder Dimension. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[45] J. Pineau, G. Gordon, and S. Thrun. Anytime Point-Based Approximations for Large POMDPs. *Jair*, 27:335-380, nov 2006.

[46] Arrasy Rahman, Elliot Fosong, Ignacio Carlucho, and Stefano V. Albrecht. Generating Teammates for Training Robust Ad Hoc Teamwork Agents via Best-Response Diversity. *Transactions on Machine Learning Research*, may 2023.

[47] João G. Ribeiro, Cassandro Martinho, Alberto Sardinha, and Francisco S. Melo. Assisting Unknown Teammates in Unknown Tasks: Ad Hoc Teamwork under Partial Observability. jan 2022.

[48] Stephane Ross and J. Andrew Bagnell. Agnostic System Identification for Model-Based Reinforcement Learning. *ArXiv:1203.1007 [cs, stat]*, jul 2012.

[49] Daniel Russo and Benjamin Van Roy. Learning to Optimize Via Posterior Sampling. feb 2014.

[50] Richard D. Smallwood and Edward J. Sondik. The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon. *Operations Research*, 21(5):1071-1088, 1973.

[51] Lauren N. Steimle, David L. Kaufman, and Brian T. Denton. Multi-model Markov decision processes. *IISE Transactions*, pages 1-16, may 2021.

[52] Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. *AAAI*, 24(1):1504-1509, jul 2010.

[53] Alexander L. Strehl and Michael L. Littman. An analysis of model-based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74(8):1309-1331, dec 2008.

[54] D. J. Strouse, Kevin R. McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with Humans without Human Data. jan 2022.

[55] Matthew E. Taylor and Peter Stone. Transfer Learning for Reinforcement Learning Domains: A Survey. *J. Mach. Learn. Res.*, 10:1633-1685, dec 2009.

[56] William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285-294, 1933.

[57] Paul Tylkin, Goran Radanovic, and David C Parkes. Learning Robust Helpful Behaviors in Two-Player Cooperative Atari Environments. 2021.

[58] D. J. White. Bayesian Decision Problems and Markov Chains. *Royal Statistical Society. Journal. Series A: General*, 132(1):106-107, jan 1969.

[59] Annie Xie, Dylan P. Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. Learning Latent Representations to Influence Multi-Agent Interaction. nov 2020.

[60] Lin Yang and Mengdi Wang. Sample-Optimal Parametric Q-Learning Using Linearly Additive Features. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6995-7004. PMLR, may 2019.

[61] Xinghu Yao, Chao Wen, Yuhui Wang, and Xiaoyang Tan. SMIX($\ lambda$): Enhancing Centralized Value Functions for Cooperative Multi-Agent Reinforcement Learning. aug 2020.

[62] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning Near Optimal Policies with Low Inherent Bellman Error. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10978-10989. PMLR, nov 2020.

[63] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. apr 2021.

[64] Tong Zhang. Feel-Good Thompson Sampling for Contextual Bandits and Reinforcement Learning. oct 2021.

[65] Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. GEC: A Unified Framework for Interactive Decision Making in MDP, POMDP, and Beyond. jun 2023.