Deep Interactive Bayesian Reinforcement Learning via Meta-Learning

Extended Abstract

Luisa Zintgraf
University of Oxford
Work done during an MSR internship

Sam Devlin
Microsoft Research

Kamil Ciosek
Work done while at MSR

Shimon Whiteson
University of Oxford

Katja Hofmann
Microsoft Research

ABSTRACT

Agents that interact with other agents often do not know a priori what the other agents' strategies are, but have to maximise their own online return *while* interacting with and learning about others. The optimal adaptive behaviour under uncertainty over the other agents' strategies w.r.t. some prior can in principle be computed using the Interactive Bayesian Reinforcement Learning framework. Unfortunately, doing so is intractable in most settings, and existing approximation methods are restricted to small tasks. To overcome this, we propose to *meta-learn* (alongside the policy) approximate belief inference by combining sequential and hierarchical VAEs. We show empirically that our approach can learn a factorised belief model that separates the other agent's permanent and temporal structure, and outperforms methods that sample from the approximate posterior or do not have this hierarchical structure. A full version of this work can be found in Zintgraf et al. [30].

KEYWORDS

Interactive Bayesian RL; Meta Learning; Variational Methods

ACM Reference Format:
Luisa Zintgraf, Sam Devlin, Kamil Ciosek, Shimon Whiteson, and Katja Hofmann. 2021. Deep Interactive Bayesian Reinforcement Learning via Meta-Learning: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 INTRODUCTION

A desirable capability of artificial agents that interact with other (human or artificial) agents is the the ability to adapt to others in an ad-hoc way, i.e., learn about the behaviour of others and adapt accordingly. For example, playing a game of soccer with a new team requires learning about each player's role and coordinating actions; driving a car through traffic requires anticipating other drivers' moves and reacting appropriately; and teaching a complex subject to a student requires adjusting the teaching method to their learning style. Many standard multi-agent reinforcement learning (RL) methods are limited in their ability to adapt to unknown other agents. In ad-hoc teamwork [27], agents trained to coordinate with each other can fail to do so when paired with unseen partners [5, 7]. This is a critical limiting factor for real-world applications.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

When faced with unknown other agents, an agent should ideally maximise *online return* incurred *while* learning about others. A principled way to study this is from the perspective of decision-making under uncertainty. An agent that acts optimally under uncertainty given a prior belief is called Bayes-optimal [13, 21], and optimally trades off exploration and exploitation. The framework for computing such agents in multi-agent settings is called Interactive Bayesian RL [IBRL; 9, 17]. This approach requires maintaining a belief over the other agents' strategies, and computing the optimal action given that belief. The policy that maximises the expected return in the resulting belief MDP [19] optimally adapts to other agents.

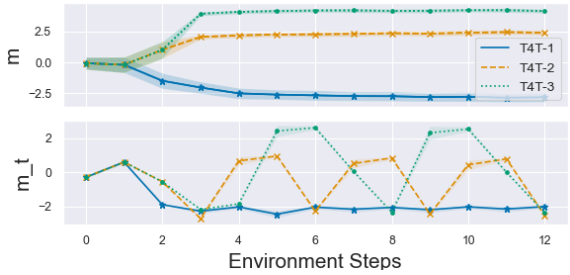
Unfortunately, computing the solution is generally intractable and existing work provides approximate solutions restricted to small environments or restrictive assumptions [8, 10, 16, 17]. In the single-agent setting, where the unknowns are the environment's transition and reward functions instead of the other agent's policies, meta-learning has recently been proposed as a scalable way to compute approximately Bayes-optimal agents [18, 23, 31].

Contribution. We argue that the IBRL framework is a useful proxy for learning adaptive policies, and propose **Me**ta **L**earning **I**nteractive **B**ayesian **A**gents (MeLIBA), a method for meta-learning approximately Bayes-optimal policies for a distribution of other agents. We leverage recent advances in agent modelling, variational inference, and meta-learning, to compute approximately Bayes-optimal agents in a general and tractable way (Sec 2). Empirically, we demonstrate that explicitly learning and conditioning on approximate beliefs over other agents' strategies can improve performance in multi-agent settings, compared to relying on samples, or using a model-free policy with memory. We show that MeLIBA learns a hierarchical latent representation of other agent types, separating the permanent and temporal internal states.

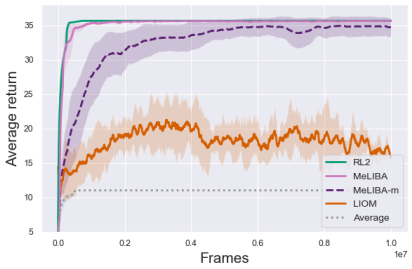
Our work differs from existing multi-agent Bayesian RL approaches [1, 6, 22, 26] in that we optimise for *Bayes*-optimal behaviour, as opposed to optimal behaviour that typically requires knowledge of the other agents' strategies. MeLIBA's latent variables can be seen as a continuous representation of agent types learned in an unsupervised way via interaction, similar to type-based modelling [2, 3, 27]. Closely related to our approach is the work of Papoudakis and Albrecht [24], who focus on partially observable settings, and approximating global from local (agent-specific) information. We allow full observability but focus on how to model agents that are non-stationary, in contrast to many existing methods that consider the other agents to be Markov [7, 14, 14, 24].

	Swerve	Straight
Swerve	(1, 1)	(0, 8)
Straight	(8, 0)	(-1, -1)

(a) Game of Chicken.



(b) Latent variables during test rollout.



(c) Learning Curves

Figure 1: (a) Payoff matrix for the Game of Chicken. (b) The latent variables of MeLIBA during 3 different games at meta-test time, when playing against the different other agents T4T-1/2/3. Top: permanent latent variable m, which separates the different agent types. Bottom: temporal latent variable m_t which keeps track of how often the agents cooperate. (c) Learning curves for MeLIBA and ablations/baselines. RL^2 and MeLIBA learn the Bayes-optimal behaviour quickly. Using only a permanent latent state (MeLIBA-m) leads to a performance drop since the other agent is not modelled appropriately. LIOM underperforms, possibly due to the small latent dimension combined with the noise from the sampling.

2 MELIBA

We propose **Me**ta **L**earning **I**nteractive **B**ayesian **A**gents (MeLIBA), a method for meta-learning approximately Bayes-optimal agents that adapt to other agents. Specifically, we jointly meta-train on a given prior distribution over other agents: (1) a belief inference network, and (2) a policy that conditions on the approximate belief.

Agent Models. We model each other agent by its own permanent latent variable (m, also called *agent character*) and a temporal latent variable (m_t, also called *mental state*). The character m does not change throughout the agent's lifetime. The mental state m_t can change in response to new observations at every timestep and allows us to model agents with non-stationary policies. This can be viewed as a probabilistic extension of Rabinowitz et al. [25], who coined the terms agent character and mental state, and used this split to model other agents in an observational setting. We instead consider an *interactive* setting, which requires us to maintain *beliefs* over the components of other agents.

Approximate Belief Inference. To perform approximate belief inference, we use a variational auto-encoder [VAE, 20] for sequential data [11] combined with a hierarchical latent structure [29]. Like in single-agent meta-learning approaches [18, 31] this VAE is trained alongside the policy, but has a more complex structure due to the possible non-stationarity of other agents.

Meta-Learning Bayes-Adaptive Policies. Given the approximate posterior, we want to learn an approximately Bayes-optimal policy. To this end, we condition our policy not only on the environment state, but also on this approximate belief over the other agents' policies. This enables approximately Bayes-optimal behaviour: the policy can take into account its uncertainty over the other agents' policies when choosing actions, and use it to trade off exploration and exploitation. In practice, we approximate the posterior using a Gaussian distribution which is fully characterised by the mean and variance of the latent distribution in the VAE. The policy is then trained using standard RL methods by conditioning on environment states and approximate beliefs. In practice, we alternate between updating the VAE, and the agent (using PPO).

3 EMPIRICAL EVALUATION

We consider a 2-player competitive matrix game, Game of Chicken (Fig 1a) [4]. Imagine two cars driving towards each other: if nobody swerves, they crash and get a penalty (−1); if they both swerve they get a medium reward (1); if only one swerves it gets a low (0) and the other a high (8) reward. We hand-code three Tit-4-Tat agents [15] which swerve if the opponent swerved 1/2/3 times in a row. We randomly sample an agent to play with for 13 repetitions. The Bayes-optimal strategy is to swerve until the other agent swerves and thereby reveals its strategy, after which it can be exploited. Fig 1c shows the performance of MeLIBA compared to other approaches. An *average* policy that cannot adapt cannot solve this task. RL^2 [12, 28] is a model-free meta-learning method with an architecture similar to MeLIBA, but with no decoder and no explicit hierarchy in the encoder. The RL loss is backpropagated through the encoder and there is no bottleneck. RL^2 learns to solve the task quickly, which is unsurprising given the simplicity of the game and since it conditions on the interaction history via the recurrent encoder. LIOM [24], where the policy receives a *sample* from the approximate posterior, and there is only a permanent latent (m), performs poorly on this task. To analyse why, we evaluate MeLIBA with *only* the fixed latent m. As Fig 1c shows, MeLIBA-m cannot solve the task given the wrong model for the other agent. However it does outperform LIOM by a large margin, indicating that sampling the latent variable, as opposed to conditioning the policy on the entire posterior, causes poor performance here.

For MeLIBA we use a latent of size 1 each for the permanent and temporal aspects. Fig 1b shows the latent mean and standard deviation of the learned beliefs when rolling out the meta-learned policy against the possible other agents. The top shows the permanent latent variable, which separates between agent types after just 1-2 timesteps. The bottom shows the same visualisation for the temporal latent variable, which counts the number of swerves.

In summary, MeLIBA builds on the IBRL framework and is a general method for meta-learning Bayes-adaptive behaviour and modelling beliefs over other agents that adapt within game. For more details and results see [30].

REFERENCES

[1] Stefano V Albrecht, Jacob W Crandall, and Subramanian Ramamoorthy. 2016. Belief and truth in hypothesised behaviours. *Artificial Intelligence* 235 (2016), 63–94.

[2] Stefano V Albrecht and Peter Stone. 2019. Reasoning about hypothetical agent behaviours and their parameters. *arXiv preprint arXiv:1906.11064* (2019).

[3] Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. 2013. Teamwork with limited knowledge of teammates. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

[4] Carl T Bergstrom and Peter Godfrey-Smith. 1998. On the evolution of behavioral heterogeneity in individuals and populations. *Biology and Philosophy* 13, 2 (1998), 205–231.

[5] Rodrigo Canaan, Xianbo Gao, Youjin Chung, Julian Togelius, Andy Nealen, and Stefan Menzel. 2020. Evaluating the Rainbow DQN Agent in Hanabi with Unseen Partners. *arXiv preprint arXiv:2004.13291* (2020).

[6] David Carmel and Shaul Markovitch. 1990. Exploration strategies for model-based learning in multi-agent systems. *Autonomous Agents and Multi-agent Systems* 2 (1990), 173–207.

[7] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the Utility of Learning about Humans for Human-AI Coordination. In *Advances in Neural Information Processing Systems*. 5175–5186.

[8] Georgios Chalkiadakis. 2007. *A Bayesian approach to multiagent reinforcement learning and coalition formation under uncertainty*. University of Toronto.

[9] Georgios Chalkiadakis and Craig Boutilier. 2003. Coordination in multiagent reinforcement learning: A bayesian approach. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. 709–716.

[10] Georgios Chalkiadakis, Edith Elkind, Evangelos Markakis, Maria Polukarov, and Nick R Jennings. 2010. Cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research* 39 (2010), 179–216.

[11] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*. 2980–2988.

[12] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. Rl2: Fast reinforcement learning via slow reinforcement learning. 2016. *arXiv preprint arXiv:1611.02779* (2016).

[13] Michael O'Gordon Duff and Andrew Barto. 2002. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. Ph.D. Dissertation. University of Massachusetts at Amherst.

[14] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. In *International Conference on Machine Learning*. 1804–1813.

[15] Shaun Hargreaves Heap and Yanis Varoufakis. 2004. *Game theory: a critical text*. Psychology Press.

[16] Trong Nghia Hoang. 2014. *New Advances on Bayesian and Decision-Theoretic Approaches for Interactive Machine Learning*. Ph.D. Dissertation. Division of the School of Computing, National University of Singapore.

[17] Trong Nghia Hoang and Kian Hsiang Low. 2013. A general framework for interacting Bayes-optimally with self-interested agents using arbitrary parametric model and model prior. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

[18] Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A Ortega, Yee Whye Teh, and Nicolas Heess. 2019. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424* (2019).

[19] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.

[20] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[21] Tom M Mitchell. 1997. Machine Learning, volume 1 of 1.

[22] John H Nachbar. 2005. Beliefs in repeated games. *Econometrica* 73, 2 (2005), 459–480.

[23] Pedro A Ortega, Jane X Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Heess, Joel Veness, Alex Pritzel, Pablo Sprechmann, et al. 2019. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030* (2019).

[24] Georgios Papoudakis and Stefano V Albrecht. 2020. Variational Autoencoders for Opponent Modeling in Multi-Agent Systems. *arXiv preprint arXiv:2001.10829* (2020).

[25] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine Theory of Mind. In *International Conference on Machine Learning*. 4218–4227.

[26] Dorsa Sadigh, S Shankar Sastry, Sanjit A Seshia, and Anca Dragan. 2016. Information gathering actions over human internal state. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 66–73.

[27] Peter Stone, Gal A Kaminka, Sarit Kraus, and Jeffrey S Rosenschein. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[28] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. 2016. Learning to reinforcement learn. In *Annual Meeting of the Cognitive Science Community (CogSci)*.

[29] Shengjia Zhao, Jiaming Song, and Stefano Ermon. 2017. Learning hierarchical features from deep generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 4091–4099.

[30] Luisa Zintgraf, Sam Devlin, Kamil Ciosek, Shimon Whiteson, and Katja Hofmann. 2021. Deep Interactive Bayesian Reinforcement Learning via Meta-Learning. *arXiv preprint arXiv:2101.03864* (2021).

[31] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. 2020. VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning. In *International Conference on Learning Representations*.