Deep Interactive Bayesian RL via Meta-Learning

Luisa Zintgraf, Sam Devlin, Kamil Ciosek, Shimon Whiteson, Katja Hofmann

Microsoft Research

UNIVERSITY OF OXFORD

Context

Question:

How can agents adapt to initially unknown other agents, while maximising *online* return?

Solution (in principle):

Interactive Bayesian RL (IBRL) [1].
Idea: Maintain belief over other agents, and compute optimal action under uncertainty.
But: IBRL is intractable for most problems.

Our key contributions: MeLIBA

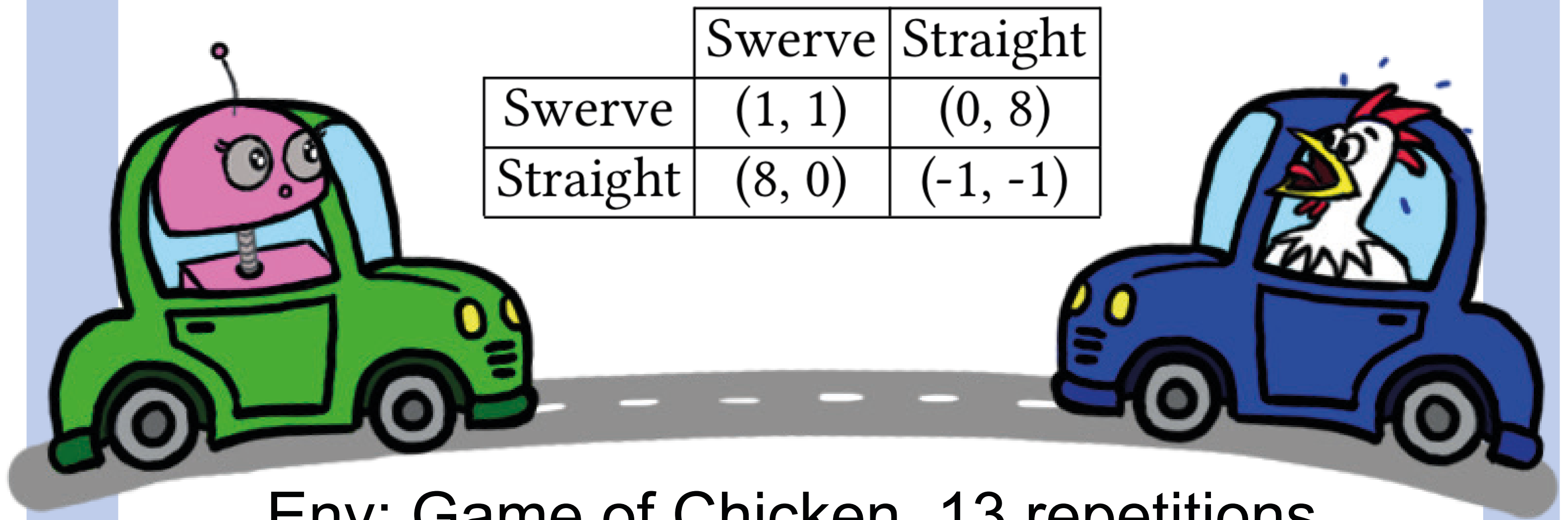
- Scaling IBRL to Deep Learning
- Learning *beliefs* over other agent's types (compared to e.g. [2])
- Learning beliefs of permanent *and temporal* aspects of other agents (compared to e.g. [3])

Future Work:

- Generate distribution of other agents (instead of hand-coding)
- Move to general POMDPs (see [3])
- Other agents that learn (during meta-training)

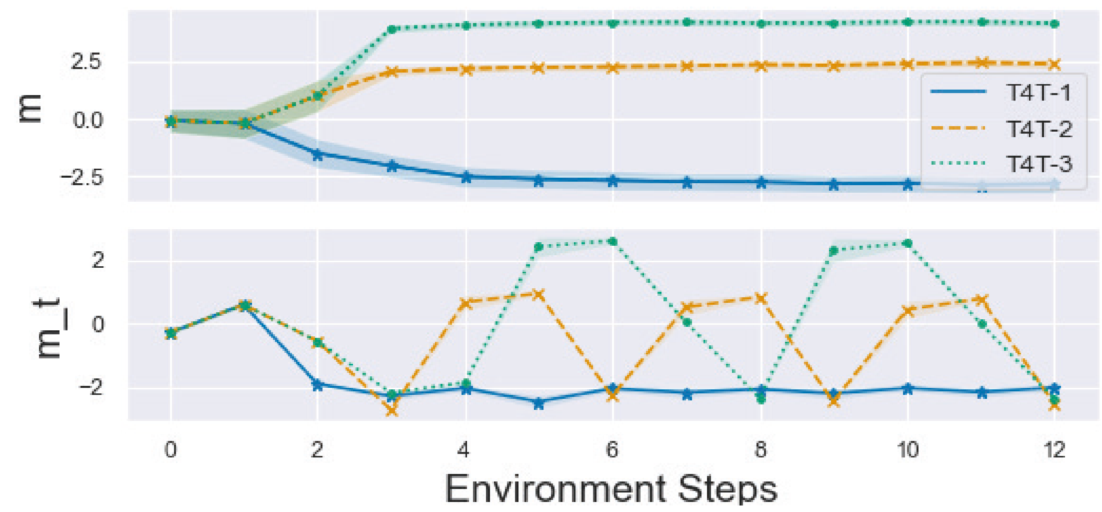
[1] Trong Nghia Hoang, Kian Hsiang Low. A general framework for inter- acting Bayes-optimally with self-interested agents using arbitrary parametric model and model prior. IJCAI 2013.
[2] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S M Ali Eslami, Matthew Botvinick. Machine Theory of Mind. ICML 2018.
[3] Georgios Papoudakis and Stefano V Albrecht. Variational Autoencoders for Opponent Modeling in Multi-Agent Systems. AAAI Workshop on RL in Games 2020.

Results

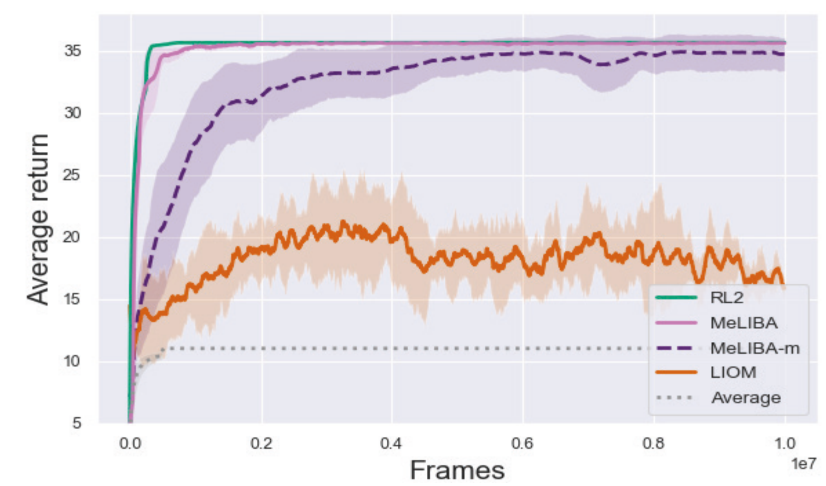


	Swerve	Straight
Swerve	(1, 1)	(0, 8)
Straight	(8, 0)	(-1, -1)

Env: Game of Chicken, 13 repetitions
Opponents: Tit-4-Tat after 1/2/3 swerves



Learned belief over other agents: separates permanent and temporal aspects



MeLIBA shows good performance compared to baselines and ablations

Method: MeLIBA

Our approach for scaling IBRL:

Meta-learn how to
(1) **infer other agents'** permanent & temporal **types** using approx. variational inference
(2) use the approximate belief for optimal **decision-making under uncertainty** over the other agents' strategies.

(1) Meta-learning belief inference:

Use a sequential hierarchical VAE trained to predict future actions of other agents, given current experience.

Separate latent for: permanent (m) and temporal (m_t) aspect of other agent.

(2) Meta-learning the policy:

Condition policy on approximate *belief*.
Trained using standard RL alongside the VAE.