Conciliating Reward-driven Models and Reward Uncertainty in Reinforcement Learning from Human Feedback

Riad Akrour[1] and Philippe Preux[1]

[1]Univ. Lille, CNRS, Inria

When and Where: 6 months (spring-summer 2024) at Scool, Inria Lille, Villeneuve d'Ascq, France.

Expected Background: Master in computer science, specialised in machine learning. This work will require the candidate to be comfortable implementing deep RL algorithms in libraries such as PyTorch.

Keywords: Deep RL, learning from human feedback, planning, model-based RL.

1 Context

In the near future, AIs will be ubiquitous in our daily lives, assisting humans on a variety of daily life tasks. Within the large spectrum of such AI assistants, those based on Reinforcement Learning (**RL**) are perhaps the most promising as they take into account the effect of each prescribed action, they continually adapt to changes in the environment and make sure the sequence of prescribed actions maximizes some objective measure of the user's needs and wants. In its current state, RL has had several achievements in various domains such as board games or robotics. However, these achievements required large teams of researchers and there is still a gap between the current state of RL methods and the aforementioned AI that solves on the fly decision making tasks for a **non-RL-expert** user. Specifically, one of the currently missing desideratum is the ability to quickly adapt to a user's preferences on complex decision making problems in a reasonable time and user feedback budget.

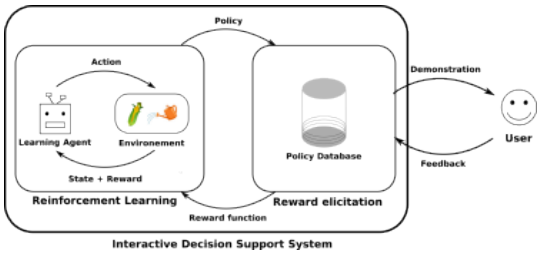

Figure 1: General framework of RL from human feedback. The AI selects solutions to demonstrate to the user, the user gives feedback on these solutions, and the AI update a model of the user's preferences and use RL to obtain new solutions to demonstrate to the use.

At Scool, Gautron [2022] have turned a high-fidelity **crop simulator** into an RL environment. In this problem, an AI advises a farmer throughout a harvesting season, deciding daily how much should the farmer water, fertilize, and so on, with a goal of striking a balance between several criteria such as yield or nitrate pollution under varying weather conditions. By running an off-the-self deep RL algorithm such as PPO [Schulman et al., 2017], it was shown in Gautron [2022] that RL can find more efficient solutions than human expert policies. However, the main drawback of the current decision support system is that it provides recommendations under a **pre-defined trade-off** between the different criteria (such as yield, pollution or work load) and can thus not adapt to the varying needs of individual farmers. An existing solution in the literature is to wrap an RL solver around a preference elicitation mechanism to allow non-RL-expert users to tune the reward function to their needs, while only interacting with the AI at a very abstract level. These methods have had a recent surge of popularity as they were shown to be

useful for training large language models [Ouyang et al., 2022]. An illustration of the RL from Human Feedback (**RLHF**) framework is given in Figure 1 and a survey can be found in Wirth et al. [2017]. In this internship, we will be interested in developing an RLHF framework to allow personalized recommendations in the context of the aforementioned precision agriculture RL problem.

2 Goals

The precision agriculture RL problem is sufficiently complex to require deep RL methods for finding competent policies. However, RLHF methods based on deep RL such as Christiano et al. [2017] or Liang et al. [2022] typically require in the order of a thousand to ten thousand user feedback which is not realistic in our setting. Prior methods that used advanced query selection mechanisms, selecting user queries that maximize some notion of information gain on the user's preferences as in Akrour et al. [2014], were significantly more **feedback efficient**, but optimizing these advanced query selection mechanisms with model-free deep RL methods might prove to be too costly in terms of **sample efficiency**. The goal of this internship is to consider planning methods with **learned models**, e.g. tree search algorithms such as MuZero [Schrittwieser et al., 2020], to find a better trade-off between feedback and sample efficiency. The open problem is that modern model-based algorithms such as MuZero learn **reward-driven** models to insure that the model captures only state information that is relevant for predicting future rewards. While important for scaling to high dimensional state spaces, it is unclear how to conciliate reward-drive model learning with the RLHF setting that iteratively refines a **distribution** over the user's **latent reward** function.

To investigate this research question, we devise four milestones that would structure this internship. Doing the first three would be appreciable, and completing all four would be outstanding.

1. Setup a simulated RLHF framework within the precision agriculture RL problem by defining at least two simulated user profiles, that provide feedback according to two different reward functions.

2. Benchmark existing RLHF literature such as Christiano et al. [2017] or Liang et al. [2022] in terms of sample and feedback efficiency.

3. Propose a mathematical framework and algorithmic solutions that would combine reward-driven model-based RL with the reward uncertainty of RLHF, and compare with existing literature.

4. Use the model-based planner to optimize advanced query mechanisms such as the Expected Utility of Selection as in Akrour et al. [2014] and compare the resulting sample and feedback efficiency with baselines of Step 2 and 3.

3 Applying

The internship will be hosted by the Inria Scool team. It will be part of the ANR project NeuRL, which funds one PhD and one research engineer position. The PhD position has already been filled. However, as part of the project, a successful intern could continue their work as a research engineer for up to one year and a half, giving them ample time to submit their work to a top machine learning conference and to look for PhD positions either at Scool our at other research labs. To apply, please send your **resume** and all available MSc **grades** to riad.akrour@inria.fr and philippe.preux@inria.fr.

References

R. Akrour, M. Schoenauer, M. Sebag, and J.-C. Souplet. Programming by feedback. In *International Conference on Machine Learning*, 2014.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 2017.

R. Gautron. *Apprentissage par renforcement pour l'aide à la conduite des cultures des petits agriculteurs des pays du Sud : vers la maîtrise des risques.* PhD thesis, 2022.

X. Liang, K. Shu, K. Lee, and P. Abbeel. Reward uncertainty for exploration in preference-based reinforcement learning. *arXiv preprint arXiv:2205.12401*, 2022.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.

C. Wirth, R. Akrour, G. Neumann, J. Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.