APRIL: Active Preference-based Reinforcement Learning

Riad Akrour **Marc Schoenauer** **Michèle Sebag**

TAO
CNRS − INRIA − Université Paris-Sud
FirstName.Name@inria.fr

Abstract

This work tackles in-situ robotics: the goal is to learn a policy while the robot operates in the real-world, with neither ground truth nor rewards.

The proposed approach is based on preference-based policy learning: Iteratively, the robot demonstrates a few policies, is informed of the expert's preferences about the demonstrated policies, constructs a utility function compatible with all expert preferences, uses it in a self-training phase, and demonstrates in the next iteration a new policy.

While in previous work, the new policy was one maximizing the current utility function, this paper uses active ranking to select the most informative policy (Viappiani and Boutilier 2010).

The challenge is the following. The policy return estimate (the expert's approximate preference function) learned from the policy parametric space, referred to as direct representation, fails to give any useful information; indeed, arbitrary small modifications of the direct policy representation can produce significantly different behaviors, and thus entail different appreciations from the expert. A behavioral policy space, referred to as indirect representation and automatically built from the sensori-motor data stream generated by the operating robot, is therefore devised and used to express the policy return estimate. In the meanwhile, active ranking criteria are classically expressed w.r.t. the explicit domain representation − here the direct policy representation. A novelty of the paper is to show how active ranking can be achieved through black-box optimization on the indirect policy representation.

Two experiments in single and two-robot settings are used to illustrate the approach.

1 Introduction

Since the early 2000s, significant advances in reinforcement learning have been obtained through using direct expert's input (inverse reinforcement learning [13], learning by imitation [6], learning by demonstration [11]), assuming the expert's ability to demonstrate quasi-optimal behaviors, and to provide an informed representation.

In 2011, two approaches based on preference learning have been proposed to learn directly a ranking-based policy [7] or a policy return estimate [2]. In the latter case, referred to as preference-based policy learning (PPL), the agent demonstrates a few policies, receives the expert's preferences about the demonstrated policies, constructs a utility function compatible with all expert preferences, uses it in a self-training phase, and demonstrates in the next iteration the policy maximizing the

current utility function. The main merit of the PPL approach is twofold. Firstly, it sidesteps the design of the reward function at the state-action level [14]; as noted by [7], this design is critical when qualitative outcomes are considered, e.g. in the cancer treatment domain. Secondly, as opposed to inverse reinforcement learning [1, 10] PPL does not require the expert to demonstrate a quasi-optimal behavior; it does not even assume that the expert knows how to solve the task (see also [15]); the expert is only required to know whether some behavior is more able to reach the goal than some other one.

PPL relies on preference learning to build the policy return estimate, an intermediate utility function used to keep the expert's burden within reasonable limits. This utility function can be thought of as a surrogate model, supporting expensive function optimization [4]. As shown by e.g. [5], active preference learning can indeed be used for interactive optimization.

Our previous work concerns the space used to learn this preference-based surrogate model. The default option is to use the input space a.k.a. direct representation, here the policy parametric space. Another option, exploiting the RL specificities and referred to as feature space or indirect representation, has also been considered. Within the latter option, the surrogate model is a weighted sum of the overall time spent in a state-action pair (i.e. the average time the policy executes a given action in a given state). The rationale for this is the following. On the one hand, this indirect representation complies with the standard RL setting under a finite time horizon, where the policy return is defined as the cumulative reward expectation in a Markov Decision Process. On the other hand, this representation is not restricted to the MDP setting, as will be shown on the cancer treatment problem [7]. Lastly, it is shown experimentally that the indirect representation is significantly more effective than the direct one [2, 3].

A second issue concerns the selection of the new policy to be demonstrated to the expert. Related approaches concerned with active optimization [9, 5, 12] proceed by generating points in the input space which maximize the expected global improvement. These approaches however do not apply when considering an indirect representation. In our previous work, an adaptive trade-off between the current utility function and an exploration term linked to the empirical success rate was used.

In the present work, we show how to take advantage of the active ranking criteria proposed by [16] to select the most informative policy to be demonstrated. In the standard setting however, both the preference estimate and the active ranking criteria are expressed on the same representation space. In our case, the preference estimate can only be learned using the indirect representation; in the meanwhile, the active ranking criteria are defined on the direct representation.

This difficulty is handled through mapping the active ranking criteria onto the indirect representation.

Two experiments in single and two-robot settings are used to illustrate the approach.

References

[1] P. Abbeel and A.Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

[2] Riad Akrour, Marc Schoenauer, and Michèle Sebag. Preference-based policy learning. In Gunopulos et al. [8], pages 12–27.

[3] Riad Akrour, Marc Schoenauer, and Michele Sebag. Preference-based reinforcement learning. In *Choice Models and Preference Learning Workshop at NIPS'11*, 2011.

[4] Andrew Booker, J. E. Dennis, Paul D. Frank, David B. Serafini, Virginia Torczon, and Michael W. Trosset. A rigorous framework for optimization of expensive functions by surrogates, 1998.

[5] E. Brochu, N. de Freitas, and A. Ghosh. Active preference learning with discrete choice data. In *Advances in Neural Information Processing Systems 20*, pages 409–416, 2008.

[6] S. Calinon, F. Guenter, and A. Billard. On Learning, Representing and Generalizing a Task in a Humanoid Robot. *IEEE transactions on systems, man and cybernetics, Part B. Special issue on robot learning by observation, demonstration and imitation*, 37(2):286–298, 2007.

[7] Weiwei Cheng, Johannes Fürnkranz, Eyke Hüllermeier, and Sang-Hyeun Park. Preference-based policy iteration: Leveraging preference learning for reinforcement learning. In Gunopulos et al. [8], pages 312–327.

[8] Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors. *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens,*

Greece, September 5-9, 2011. Proceedings, Part I, volume 6911 of *Lecture Notes in Computer Science*. Springer, 2011.

[9] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[10] J. Zico Kolter, Pieter Abbeel, and Andrew Y. Ng. Hierarchical apprenticeship learning with application to quadruped locomotion. In *NIPS*. MIT Press, 2007.

[11] G. Konidaris, S. Kuindersma, A. Barto, and R. Grupen. Constructing skill trees for reinforcement learning agents from demonstration trajectories. In *NIPS*, pages 1162–1170. 2010.

[12] Rémi Munos and Andrew W. Moore. Rates of convergence for variable resolution schemes in optimal control. In Pat Langley, editor, *ICML*, pages 647–654. Morgan Kaufmann, 2000.

[13] A.Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In P. Langley, editor, *Proc. of the Seventeenth International Conference on Machine Learning (ICML-00)*, pages 663–670. Morgan Kaufmann, 2000.

[14] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 1998.

[15] U. Syed and R. Schapire. A game-theoretic approach to apprenticeship learning. In *NIPS*, pages 1449–1456, 2008.

[16] Paolo Viappiani and Craig Boutilier. Optimal bayesian recommendation sets and myopically optimal choice query sets. In *NIPS*, pages 2352–2360, 2010.