Preference-based Evolutionary Direct Policy Search

Róbert Busa-Fekete[*†] *Balázs Szörényi*[†‡] *Paul Weng*[§] *Weiwei Cheng*[*]
Eyke Hüllermeier[*]

Preference-based reinforcement learning (PBRL) is a novel research direction combining reinforcement learning (RL) and preference learning [3]. It aims at extending existing RL methods so as to make them amenable to training information and external feedback more general than numerical rewards, which are often difficult to obtain or expensive to compute. For example, what is the cost of a patient's death in a medical treatment?

Akrour et al. [1] and Cheng et al. [2] tackle the problem of learning policies solely on the basis of pairwise comparisons between trajectories, suggesting that one system behavior is preferred to another one but without committing to precise numerical rewards. Building on novel methods for preference learning, this is accomplished by providing the RL agent with qualitative policy models, such as ranking functions. More specifically, Cheng at al. train a model that ranks actions given state, using a method called *label ranking*. Their approach generalizes classification-based approximate policy iteration [6]. Instead of ranking actions given states, Akrour et al. exploit preferences on trajectories in order to learn a model that ranks complete policies.

In this study, we present a preference-based extension of *evolutionary direct policy search* (EDPS) as proposed by Heidrich-Meisner and Igel [5]. As a direct policy search method, it shares commonalities with [1], but also differs in several respects. In particular, their approach is arguably more specialized and tailored for applications in robotics, in which a user interacts with the learner in an iterative process. Moreover, policy search is not performed in a parametrized policy space directly but in a *feature space* capturing important background knowledge about the task to be solved.

EDPS casts policy learning as a search problem in a parametric policy space, where the function to be optimized is a performance measure like expected total reward, and evolution strategies (ES) such as CMA-ES [4] are used as optimizers. Moreover, since the evaluation of a policy can only be done approximately, namely in terms of a finite number of *rollouts*, the authors make use of *racing algorithms* to control this number in an adaptive manner. These algorithms return a sufficiently reliable ranking over the current set of policies (candidate solutions), which is then used by the ES for updating its parameters and population. A key idea of our approach is to extend EDPS by replacing the *value-based* racing algorithm with a *preference-based* one. Correspondingly, the development of a preference-based racing algorithm can be seen as a core contribution of our approach.

Value-based vs. preference-based racing setup

In this subsection we describe value-based racing setup with respect to preference-based one. We start our description with the original value-based setup introduced in [7, 8]. Let $X_1, \ldots, X_K$ be random variables with respective (unknown) distribution functions $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_K}$. These random variables, subsequently also called *options*, are supposed to have finite expected values $\mu_i = \int x d\,\mathbb{P}_{X_i}(x)$. The racing task consists of selecting, with a predefined confidence $1 - \delta$, a κ-sized subset of the K options with highest expectations. In other words, one seeks a set $I \subseteq [K]$ of cardinality κ

[*]Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str., 35032 Marburg, Germany
[†]Research Group on Artificial Intelligence, Hungarian Academy of Sciences and University of Szeged, Hungary
[‡]INRIA Lille - Nord Europe, SequeL project, 40 avenue Halley, 59650 Villeneuve d'Ascq, France
[§]Laboratory of Computer Science of Paris 6, University Pierre and Marie Curie, 4 place Jussieu, 75005 Paris, France

maximizing $\sum_{i \in I} \mu_i$, which is equivalent to the following optimization problem:

$$\sum_{i \in I} \sum_{j \neq i} \mathbb{I}\{\mu_j < \mu_i\} \longrightarrow \max_{I \subseteq [K]:\, |I| = \kappa} , \tag{1}$$

The Hoeffding race (HR) algorithm [7, 8] is an adaptive sampling method that makes use of the Hoeffding bound to construct confidence intervals for the empirical mean estimates of the options. Then, in the case of non-overlapping confidence intervals, some options can be eliminated from further sampling.

Our preference-based racing setup assumes K random variables $X_1, \ldots, X_K$ with distributions $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_K}$, respectively, and these random variables take values in a partially ordered set $(\Omega, \prec)$. Obviously, the value-based racing setup is a special case, with $\Omega = \mathbb{R}$ and $\prec$ reduced to the standard $<$ relation on the reals (comparing rollouts in terms of their rewards). The preference relation we consider here is defined as $X \ll Y$ if and only if $\mathbb{P}(Y \prec X) < \mathbb{P}(X \prec Y)$, where $\mathbb{P}(Y \prec X)$ denotes the probability that the realization of X is preferred (with respect to $\prec$) to the realization of Y.

The goal of our preference-based racing (PBR) algorithm is to find the best κ random variables with respect to the decision model $\ll$ introduced above. This leads to the following optimization task:

$$\sum_{i \in I} \sum_{j \neq i} \mathbb{I}\{X_j \ll X_i\} \longrightarrow \max_{I \subseteq [K]:\, |I| = \kappa} \tag{2}$$

In our talk we will present an algorithm that solves this optimization problem with high probability and we test it in preference-based reinforcement learning. Our synthetic experiments on medical treatment design are promising and demonstrates the versatality of our preference-based EDPS approach.

References

[1] R. Akrour, M. Schoenauer, and M. Sebag. Preference-based policy learning. In *Proceedings ECMLPKDD 2011*, pages 12–27, 2011.

[2] W. Cheng, J. Fürnkranz, E. Hüllermeier, and S.H. Park. Preference-based policy iteration: Leveraging preference learning for reinforcement learning. In *Proceedings ECMLPKDD 2011,*, pages 414–429, 2011.

[3] J. Fürnkranz and E. Hüllermeier, editors. *Preference Learning*. Springer-Verlag, 2011.

[4] N. Hansen and S. Kern. Evaluating the CMA evolution strategy on multimodal test functions. In *Parallel Problem Solving from Nature-PPSN VIII*, pages 282–291, 2004.

[5] V. Heidrich-Meisner and C. Igel. Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search. In *Proceedings of the 26th ICML*, pages 401–408, 2009.

[6] M. Lagoudakis and R. Parr. Reinforcement learning as classification: Leveraging modern classifiers. In *Proceedings of the 20th ICML*, pages 424–431, 2003.

[7] O. Maron and A.W. Moore. Hoeffding races: accelerating model selection search for classification and function approximation. In *NIPS*, pages 59–66, 1994.

[8] O. Maron and A.W. Moore. The racing algorithm: Model selection for lazy learners. *Artificial Intelligence Review*, 5(1):193–225, 1997.