Integrity Preserving Policy Learning

PhD Topic, 2011-2014

co-supervisors
Michele Sebag, CNRS & U. Paris-Sud
François Laviolette, U. Laval Qubec

June 19, 2011

Robotic policy learning *in-situ* requires specific strategies in order to preserve the robot physical integrity during the exploration of its environment. The goal of the PhD is to propose such strategies, define and establish some PAC guarantees about the incurred risk and experimentally investigate their performance.

Former approaches to integrity-preserving policy learning assume a Lipschitz environment (transition and reward functions) and consider an archive of agent traces (Fonteneau PhD, 2011). Lipschitz properties enable to provide lower bounds on the return performance of new policies. A cautious RL approach is designed, called CGRL, by maximizing such lower bounds. The proposed subject aims at both getting rid of the Lipschitz assumption and integrating some exploration within the search (whereas CGRL exploration is restricted to combining the trace fragments).

Two approaches might be considered, inspired from developmental robotics. The idea is to gradually modify the setting, and/or to maintain some knowledge about the safe behaviors in the environment. In the first case, the robot grows from a "child" to an "adult" status, granted that a child has a limited range of actions compared to an adult and thus incurs less risk. In the second case, the robot builds a *deja-vu* model of the (state,action) pairs which have been visited; the action strength, e.g. the move speed, decreases inversely proportionally to the *deja-vu* feeling, up to stopping when the robot arrives in a completely unknown state.

Formally, the first approach might consider a family of Markov Decision Process MDP$_\gamma = (\mathcal{S}, \mathcal{A}, T_\gamma, R)$ with $\mathcal{S}$ the set of states, $\mathcal{A}$ the set of actions, T_γ the transition model when the action "strength" is limited by $\gamma \in]0,1]$, and R the reward function. Given the optimal policy π^*_γ in MDP$_\gamma$, the goal is to find $\pi^*_{\gamma+d\gamma}$ and limit the exploration as much as possible (warm restart). Extending the principled use of a meta-MDP to identify the differences between two MDPs in the general case (Zhioua et al. 2009) on the one hand and relying on on π^*_γ on the other hand, one will focus the exploration on the state/action steps which are visited by π^*_γ and where MDP$_\gamma$ and MDP$_{\gamma+d\gamma}$ most differ.

The second approach maintains an archive $\mathcal{E}$ of the (state, action) which have been visited together with the cumulative reward gathered on the trajectory when they were visited (the max cumulative reward in the case they were visited in several trajectories). The expert's feedback (Akrour et al. 2011) might be used instead of the cumulative reward.

The support of the visited (state,action) pairs is learned using e.g. a One class SVM, and used to adjust locally the slowing down factor γ mentioned in the previous approach. A critical issue is to avoid increasing the support of the visited pairs after each simulation, which would end up in considering that the neighborhood of every visited point is safe, and finally that every (state, action) is safe. The points included in $\mathcal{E}$ need thus be filtered out, and the idea is to use the information given by the cumulative reward during the filtering step. Ultimately, the idea is to dynamically prune *a priori* some branches of the exploration tree/graph by using the *deja-vu* model.

References

- R. Fonteneau, S. A. Murphy, L. Wehenkel, D. Ernst: A Cautious Approach to Generalization in Reinforcement Learning. ICAART (1) 2010: 64-73

- Zhioua, S., Precup, D., Laviolette, F., Desharnais, J.: Learning the Difference between Partially Observable Dynamical Systems. In ECML/PKDD (2)(2009) 664-677

- Akrour R., Schoenauer M., Sebag M.: Preference-based Policy Learning. In ECML/PKDD 2011, to appear.