Self-Organizing Exploration in Reinforcement Learning

Simón Smith Bize and J. Michael Herrmann
Institute of Perception, Action and Behaviour, School of Informatics
University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, U.K.
artificialsimon@ed.ac.uk, j.michael.herrmann@gmail.com

Abstract

In order to solve dynamical optimization problems, reinforcement learning (RL) constructs a value function and/or a control policy. Practically, this leads, however, to suboptimal behavior: In high dimensions, the exploration of the state space becomes a time-consuming task or, in the continuous case, gradient-based RL is prone to local optima, because an RL-controlled robot is biased towards paths that lead to optimal future reward and necessarily reduces the exploration of other regions in the state space. The value gradient may even be unknown such that probing actions are necessary. Nevertheless, high-frequency probing seems appropriate for an autonomous agent that is not able to go back to square one and restart its policy. In addition, the set-up of the probing actions requires some domain knowledge and becomes cumbersome in high dimensions.

In the present contribution, we propose a self-organizing (SO) control strategy for exploring the state space around promising regions instead of a random or designed exploration strategy. The SO algorithm searches for optimal controllability and optimal observability and serves the main RL controller as a subcontroller. While the main controller generates motor signals given the actual states and improves these by RL, the SO controller generates exploratory motors signals that are used by the RL controller as exploratory actions in order to update the parameters values in an actor-critic configuration. The SO controller is updated based on the error with respect to the predicted reward. We show thus an integration of two approaches to the unsupervised generation of behaviour in robots where the interaction is based on an objective function that maximizes the sensitivity of the learning systems with respect to mismatches in the value function while simultaneously an RL component aims at maximizing the future reward.

We have tested our approach with a pendulum swing-up task and with a walking hexapod in order to assess tasks of different complexity. Our results show that (1) the exploration induced by the SO controller may counteract the reward maximization in an optimally tuned low-dimensional task, while (2) the SO controller seems to aid the learning process by guiding the exploration in a high dimensional task, and that (3) the variable coherency of the action modulation in the SO controller improves the capability of the algorithm to escape local minima and flat regions of the goal function. This approach implies that an SO controller used as a probing signal for an RL controller can reduce the time needed for fitting the policy by visiting such states that are more sensitive and by exploring unvisited states that are off the preliminarily optimal paths. Restarts of the state is often necessary in RL with random exploration, but it is not required in the present self-explorative variant. If a stable performance is reached at any local or global optimum the sensitivity of the SO controller increases until the state of the systems escapes from the stationary behaviour. We will also report on current work including the comparison of the quality and frequency of the visited states as well as systematic assessment of the scaling properties which are promising as we have shown recently in the context of guided self-organization.

Figure 1: The hexapod robot implemented in the LpzRobots simulator

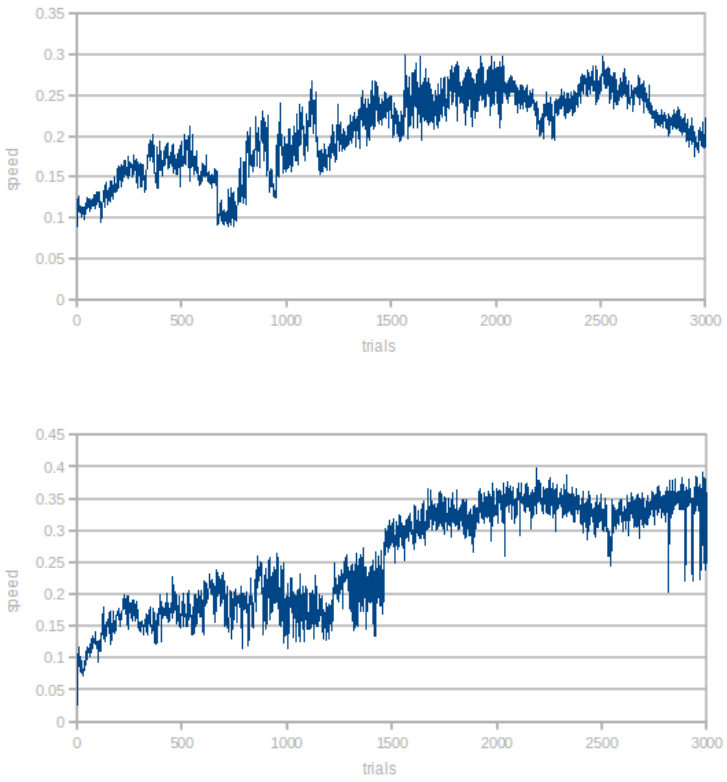

Figure 2: Results of the experiments on the hexapod robot with speed as reward. On the left RL controller, on the right SO controller was added as probing signal (note the different scales)