Direct Policy Search (DPS) denotes a class of Reinforcement Learning algorithms that allow to learn directly optimal policies without approximating their value functions. DPS has been particularly successful in continuous, noisy, and potentially non-markovian real-world domains. Unfortunately, DPS requires often a large number of interactions with the environment to learn a good policy. In this poster, we present a novel combination of DPS with model-based learning that significantly improves sample-efficiency.

Abstract

Model-based Direct Policy Search

Benchmark: Mountain Car

Experimental Setup

Results

References

Model Learner

- Remember all transitions $T = \{(s_0, a_0, s_1), \ldots, (s_n, a_n, s_{n+1})\}$, and partition for actions $T_a = \{(s_i, r_i, s'_i) \mid (s_i, r_i, s'_i) \in T \land a_i = a\}$
- Sample instances in state $s$ with probability $P_{sa} : T_a \rightarrow [0, 1]$, $P_{sa}(s_i, r_i, s'_i) = w_{sa}/\sum_{(s_j, r_j, s'_j)_{T_a}} w_{sa}$, with $w_{sa} = \exp\left(-\frac{(s_i - s'_i)^2}{b}\right)$
- Sample successor state $s'$ and reward $r$ for applying action $a$ in state $s$ as: $s' = s + (s' - s)$ and $r = r$
- $R_{max}$-based exploration: $r = \begin{cases} R_{max} & \text{if} \sum_{(s_j, r_j, s'_j)_{T_a}} w_{sa} < T_{expl} \\ r & \text{else} \end{cases}$

Direct Policy Search

- Given a class of policies $\Pi$ parametrized by a vector $\theta$
- Use metaheuristic to search in parameter space for a vector $\theta^*$ that maximizes $F(\pi_{\theta^*})$
- MBDPS: Compute $F(\pi_{\theta^*})$ solely based on trajectories sampled from internal model

Evaluator

- For a given policy and state distribution, use accumulated reward which is obtained when state transitions and rewards are sampled from model and actions are chosen based on policy as the policy’s fitness estimate.

Contact:
Jan Hendrik Metzen
DFKI Bremen & University Bremen
Robotics Innovation Center
Director: Prof Dr Frank Kirchner
www.dfki.de/robotics
jhm@informatik.uni-bremen.de

Jan Hendrik Metzen and Frank Kirchner

References

Nikolaus Hansen and Andreas Ostermeier.
Completely derandomized self-adaptation in evolution strategies.
Nicholas K. Jong and Peter Stone.
Model-based function approximation in reinforcement learning.
Jan Hendrik Metzen and Mark Edgington.
Maja Machine Learning Framework.