Meta Reinforcement Learning

The Big Picture

Marius Lindauer

tnt

Leibniz Universität Hannover

Automated Machine Learning Hannover

AutoML.org

RL Problem Setting

- Definition of MDP (S, A, P, R, γ)
  - S is a (finite) set of Markov states $s \in S$
  - A is a (finite) set of actions $a \in A$
  - P is dynamics/transition model for each action, that specifies $P(s_{t+1} = s' \mid s_t = s, a_t = a)$
  - R is a reward function $R(s_t = s, a_t = a) = \mathbb{E}[r_r \mid s_t = s, a_t = a]$
    - Sometimes R is also defined based on (s) or on (s, a, s')
  - Discount factor $\gamma \in [0, 1]$

- Task: Compute the optimal policy

$$\pi^*(s) \in \arg\max_{\pi} V^{\pi}(s)$$

Automated Machine Learning Hannover

Can We Generalize beyond a given MDP?

- What happens if the environment changes? (non-stationary environments)
  - Can we efficiently adapt our policy to changed transitions or reward functions?
- After a human player learned how to play Super Mario Bros in the first levels, they will also play fairly well the upcoming levels.
- However, an RL agent potentially will fail.

$\rightsquigarrow$ Strong limitations regarding the applications of a trained agent

Three Main Ideas

1. Can we use easy environments to learn how to behave in hard environments?

Three Main Ideas

1. Can we use easy environments to learn how to behave in hard environments?
2. Can we train a policy that is easily adaptable to new environments?

Three Main Ideas

1. Can we use easy environments to learn how to behave in hard environments?
2. Can we train a policy that is easily adaptable to new environments?
3. Can we find better training dynamics across a set of environments?

Three Main Ideas

1. Can we use easy environments to learn how to behave in hard environments?
2. Can we train a policy that is easily adaptable to new environments?
3. Can we find better training dynamics across a set of environments?
4. Can we train a policy that generalizes to new environments without any new training?

- Assumption: We sample our environments i.i.d. from a fixed distribution
  - Similar to the assumption in supervised learning, but on a meta-level
  - Training environments to train our agent on and test environments to check how well it performs.
  - We might have control how we sample from this distribution; we might don't.