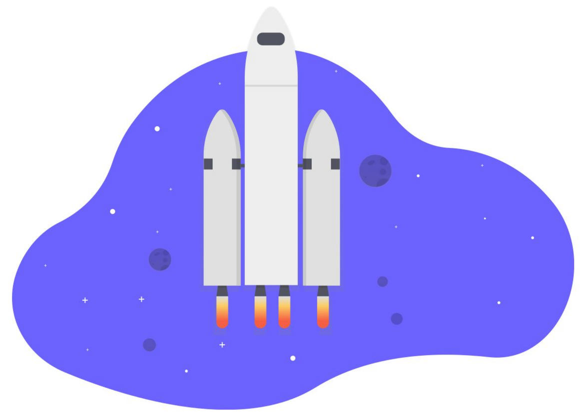
Self-Paced Context Evaluation for Contextual Reinforcement Learning



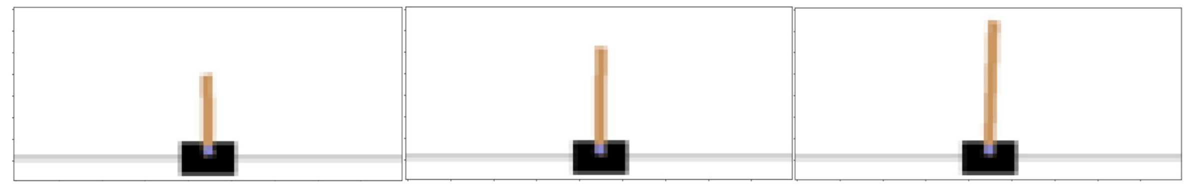
Theresa Eimer[1], André Biedenkapp[2], Frank Hutter[2 3], Marius Lindauer[1]

[1]Leibniz University Hannover | [2]Albert-Ludwigs University Freiburg | [3]Bosch Center for Artificial Intelligence

Leibniz Universität Hannover

Contextual Reinforcement Learning

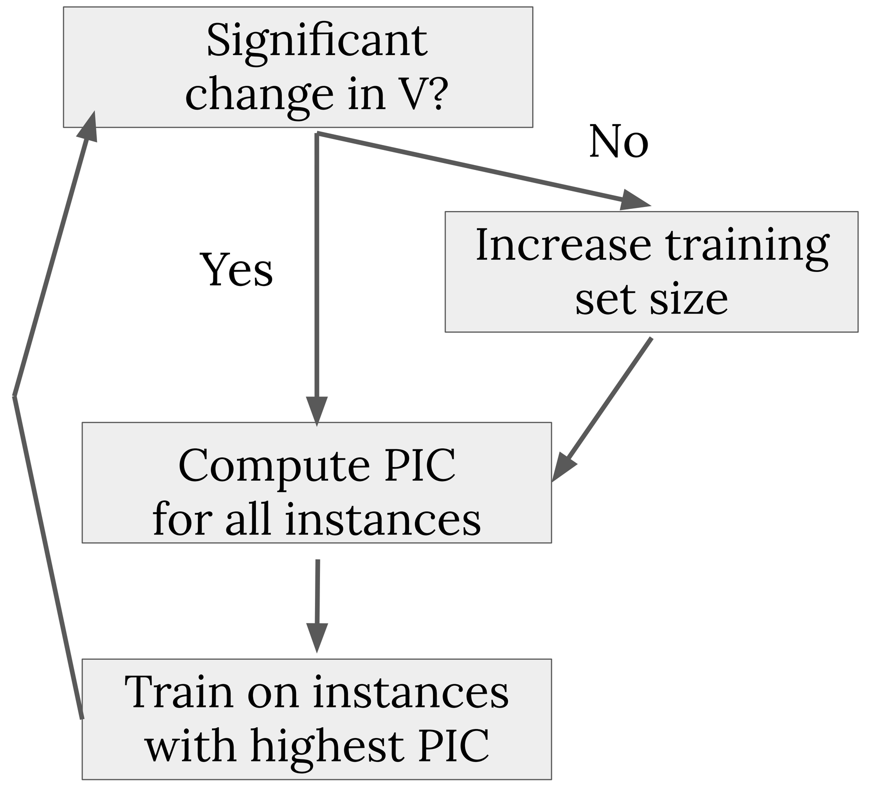
- Extending RL with task instances
- Each instance is defined by a context, e.g. the pole length in CartPole
- Requires generalization to solve



- A cMDP $\mathbf{M_I} := \{\mathbf{M_i}\}_{\mathbf{i} \in \mathbf{I}}$ consists of an MDP **M** for each instance **i** of a given instance set **I**
- Between different $\mathbf{M_i}$ actions **A** and state space **S** stay the same
- Transition dynamics **T** and reward function **R** can vary depending on the instance context
- We assume the agent is given the context during training

SPaCE in a Nutshell

- Creating instance curricula using the agent's value function V
- Change in V as proxy for agent capability (PIC)
- Difficulty rating: difference in V between training steps
- Start training on few instances and increase over time
- New instances are used whenever V converges



Algorithm 1: SPaCE curriculum generation

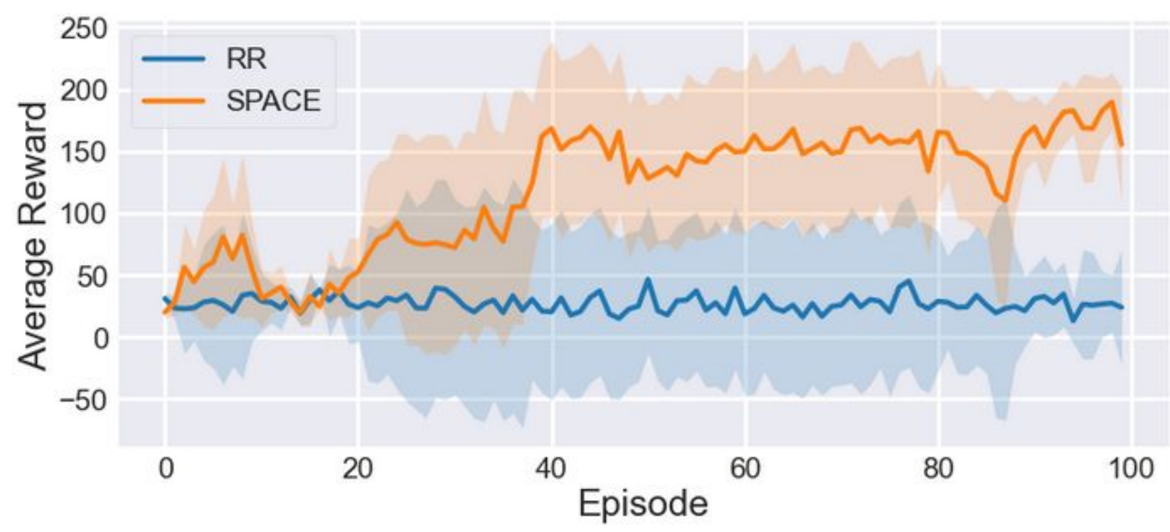
Data: policy π, value function V, Instance set $\mathcal{I}$, threshold η, step size κ, #iterations T

$S, t := 0$
$V_0 := 0$
$\mathcal{I}_{curr} := \{i\}$ with i randomly sampled from $\mathcal{I}$
for $t = 1...T$ **do**
  $\pi, V_t^\pi := update(\pi, V_{t-1}^\pi, \mathcal{I}_{curr})$
  $V_t^\pi := \frac{1}{|\mathcal{I}_{curr}|} \sum_{i \in \mathcal{I}_{curr}} |V_t^\pi(s_0, c_i)|$
  if $V_t^\pi \in [(1-\eta)V_{t-1}^\pi, (1+\eta)V_{t-1}^\pi]$ **then**
    // Increase set size
    $S := S + \kappa$
  // Choose next instance set
  forall $i \in \mathcal{I}$ **do**
    $d_t(i) := V_t^\pi(s_0, c_i) - V_{t-1}^\pi(s_0, c_i)$
  $\mathcal{I}_{curr} := S$ instances with highest $d_t(i)$
  $t := t + 1$

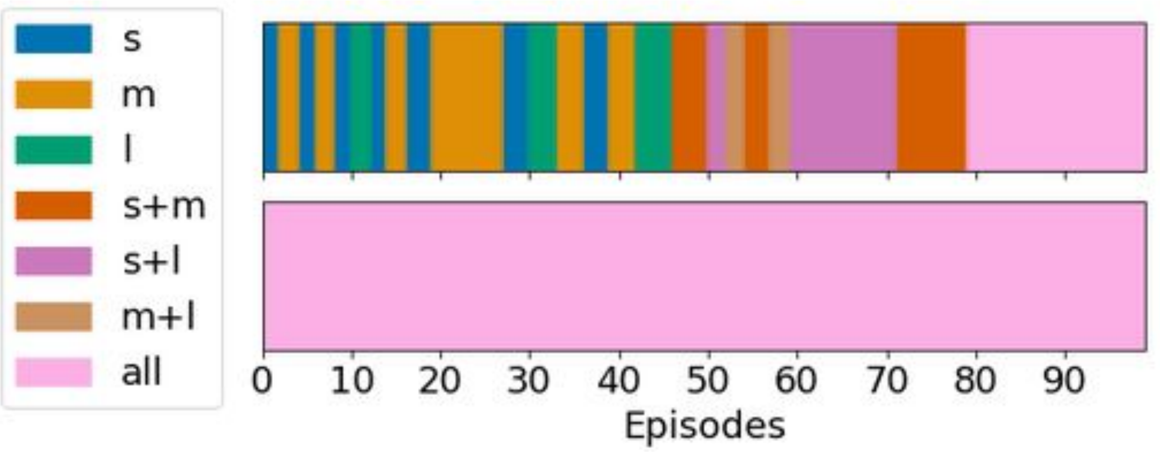
AutoML.org

Experimental Results

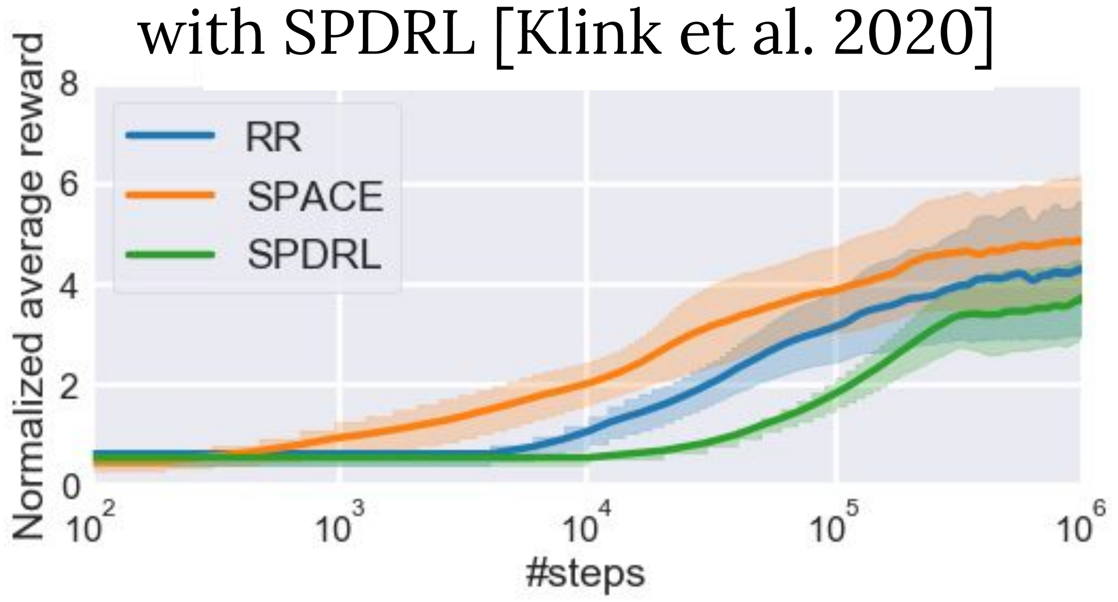
Contextual **CartPole** Test Performance



Contextual **CartPole** Curriculum



PointMass Test Performance with SPDRL [Klink et al. 2020]



Main Take-Aways:

- Better generalization performance as well as better sample efficiency during training
- Difficulty progression in curricula is not always linear, but successfully goes from easy to difficult

Why use SPaCE?

	Domain knowledge independent	Improved sample efficiency	Better overall generalization
Environment Evolution [Wang et al. 2019]	✗	✓	✓
Curricula through Self-Play [Sukhbaatar et al. 2018]	✓	✓	✗
Student-Teacher approaches [Matiisen et al. 2019]	✗	✓	✓
Difficulty appropriate instance sampling [Klink et al. 2020]	✗	✓	✓
SPaCE (ours)	✓	✓	✓

UNI FREIBURG

Blog Post

@AutoML_org