An Automated Measure of Similarity for Transfer in Reinforcement Learning

Haitham Bou Ammar, Eric Eaton, Matthew Taylor,
Decebal Mocanu, Kurt Driessens, Karl Tuyls, and Gerhard Weiss

Contribution: A data-driven similarity measure between reinforcement learning tasks.

Motivation

Transfer Learning aims to improve learning times on a new target task by reusing knowledge from previously learned source task(s). In transfer, the performance of any algorithm depends on the choice of the source and target tasks. Here, we present a data-driven similarity measure used to choose source task(s).

Markov Decision Processes

Tasks are modelled as Markov Decision Processes (MDPs). An MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ with:

$\mathcal{S}$: state space

$\mathcal{A}$: action space

γ : discount factor

$\mathcal{P}$: transition probability

$\mathcal{R}$: reward function

RBDist Similarity Measure

Intuition: If two tasks are similar, then a restricted Boltzmann Machine (RBM) trained on samples from the first task should reconstruct samples from the other task. The distance is measured using two phases:

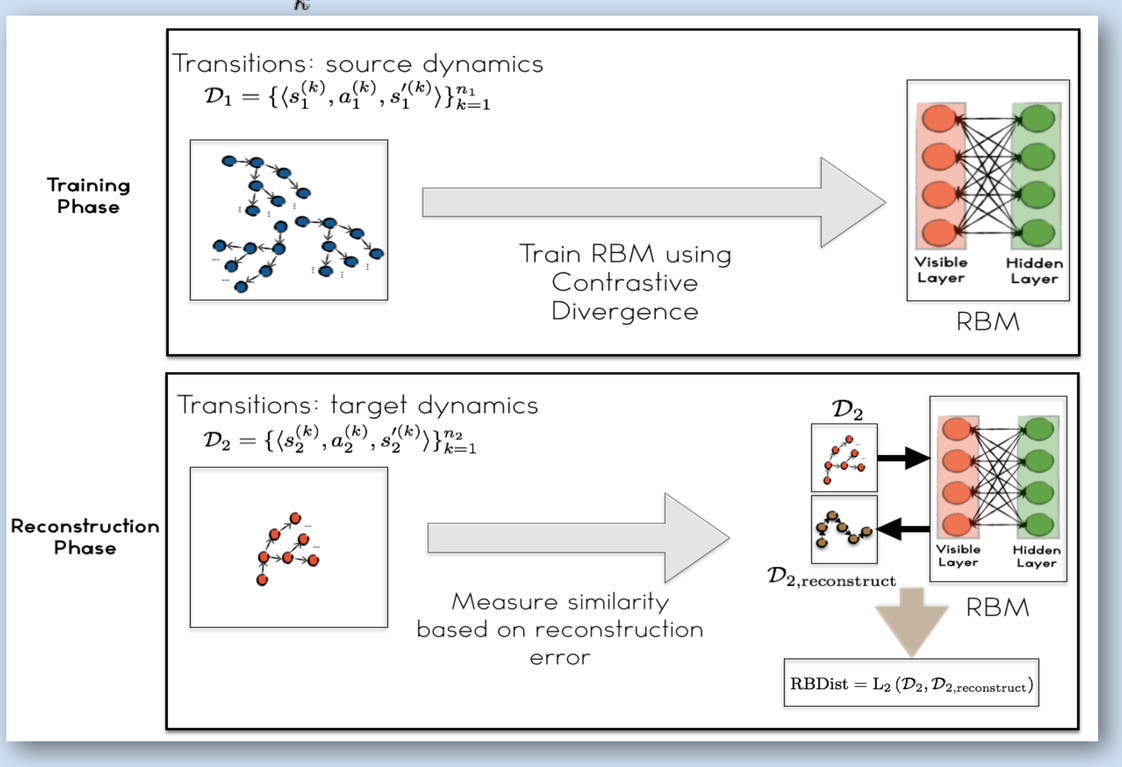
Training Phase: Using ***source samples,*** train an RBM by contrastive divergence.

Reconstruction Phase: Reconstruct ***target samples*** by sampling the visible layer (having conditionally independent visible units)

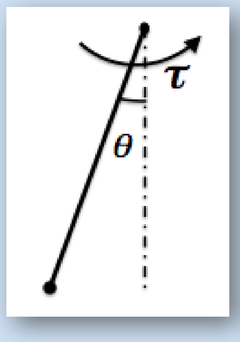
$$p\left(v|h,w\right) = \prod_{i=1}^{n_v} \mathcal{N}\left(\sum_f w_{if} h_f + b_i\right)$$

Measure similarity: using the Euclidean measure between real samples and reconstructed ones

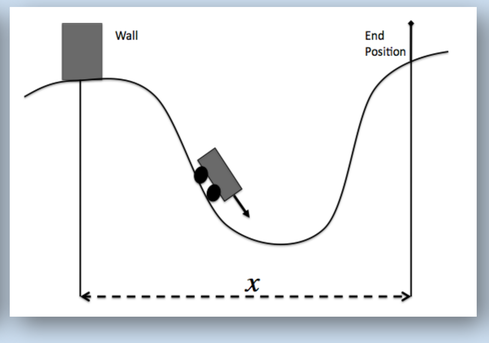
$$E = \frac{1}{n} \sum_k \mathrm{L}_2 \left(\langle s_2^k, a_2^k, s_2^{\prime(k)} \rangle_0, \langle s_2^k, a_2^k, s_2^{\prime(k)} \rangle_1 \right)$$



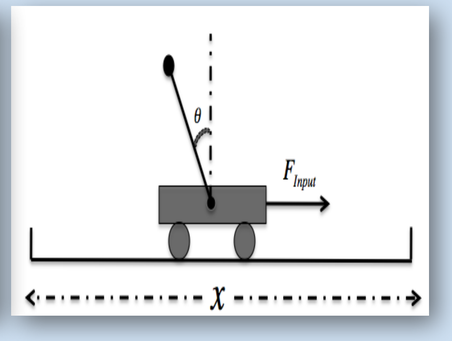
Experimental Domains & Benchmarks



Inverted Pendulum

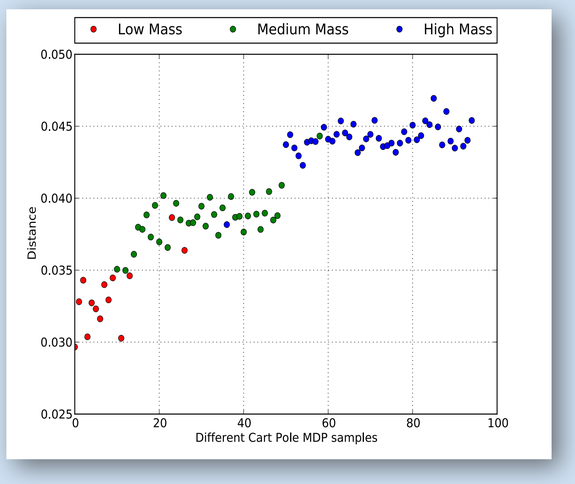


Mountain Car

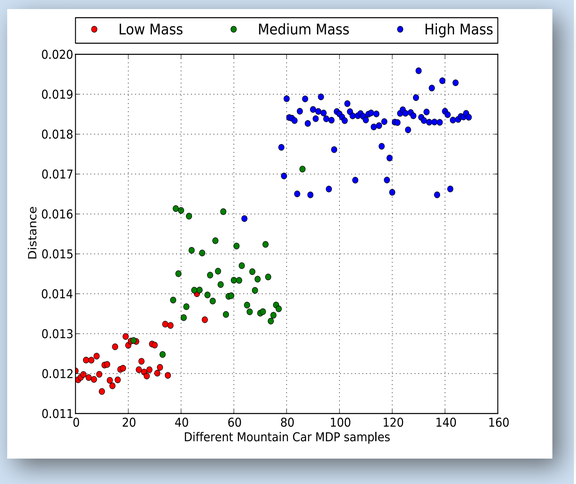


Cart Pole

Dynamical Phase Discovery



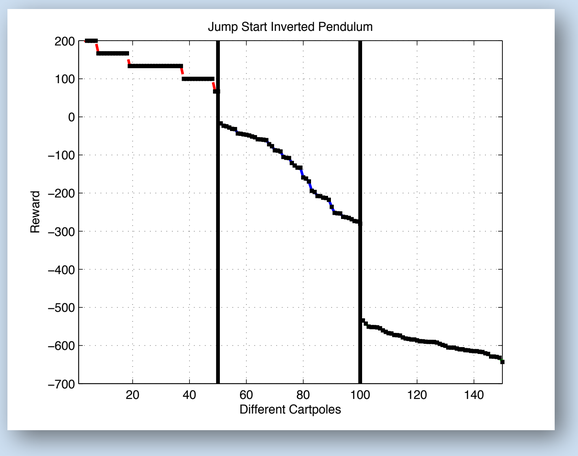
RBDist results for clustering Cart Pole systems



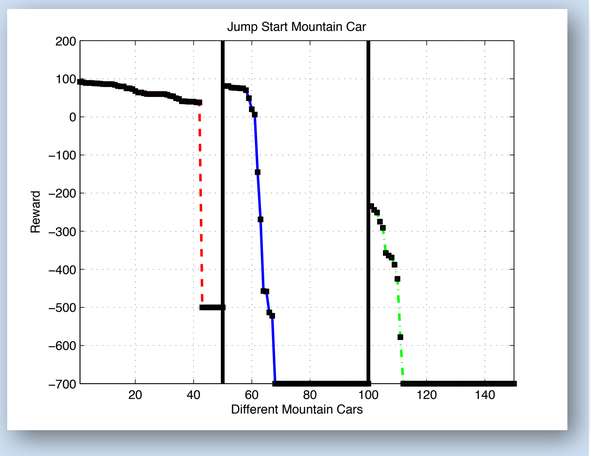
RBDist results for clustering Mountain Car systems

RBDist can automatically discover tasks' dynamical phases

Transfer Correlation



Jump-Start correlation as a function of RBDist on Cart Pole systems



Jump-Start correlation as a function of RBDist on Mountain Car systems

RBDist correlates with initial performance on target tasks

Future Work

- Extend RBDist to support transfer between different domain tasks
- Assess the effect of RBDist on other transfer criteria (e.g., asymptotic performance, time to threshold, etc.)

Correspondence to:
Haitham Bou Ammar (haithamb@cis.upenn.edu)
Eric Eaton (eeaton@cis.upenn.edu)