An Automated Measure of MDP Similarity for Transfer in Reinforcement Learning

Haitham Bou Ammar

Eric Eaton

Penn Engineering

Matthew Taylor

WASHINGTON STATE UNIVERSITY

Decebal Mocanu

TU/e Technische Universiteit Eindhoven University of Technology

Kurt Driessens

Karl Tuyls

Maastricht University

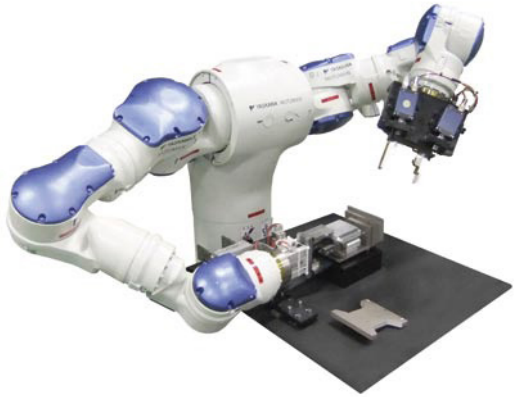
Gerhard Weiss

Maastricht University

UNIVERSITY OF LIVERPOOL

Introduction

Reinforcement learning (RL) is a key technique for learning through interaction with the environment



Problem Definition:

RL problems are formalized as **Markov Decision Processes (MDPs):** $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

$\mathcal{S}$: State Space

$\mathcal{A}$: Action Space

γ : Discount Factor

$\mathcal{P}$: Transition Probability

$\mathcal{R}$: Reward Function

Goal

Learn optimal policy by maximizing

$$Q(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_t\right]$$

Motivation

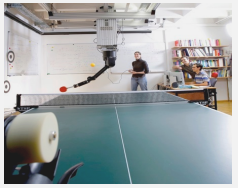
Problem

Reinforcement learners are
slow to learn

Possible Solution

Reuse knowledge
from other sources

Impressive Results


- Learning from Demonstration
- Transfer Learning

Transfer Learning

Pool of source tasks from same domain

New target task

Questions to answer:

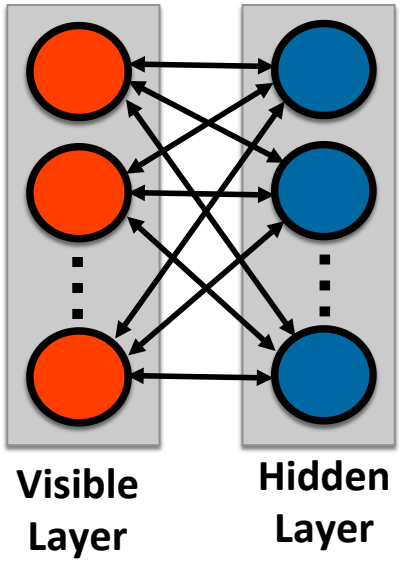
1. How to transfer?
   - → lots of approaches
2. What to transfer?
   - → lots of approaches
3. When to transfer?
   - → Less progress has been achieved → Needs a task similarity measure

RBDist: Similarity Measure Between MDPs

RBDist: Similarity Measure Between MDPs

Our measure is based on
Restricted Boltzmann Machines (RBMs):

- Set of visible units $\mathcal{V} = \{v^{(1)}, \ldots, v^{(n_v)}\}$
- Set of hidden units $\mathcal{H} = \{h^{(1)}, \ldots, h^{(n_h)}\}$



RBM Energy Function

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i,j} v^{(i)} h^{(j)} w^{(i,j)} - \sum_{i} v^{(i)} a^{(i)} - \sum_{j} h^{(j)} b^{(j)}$$

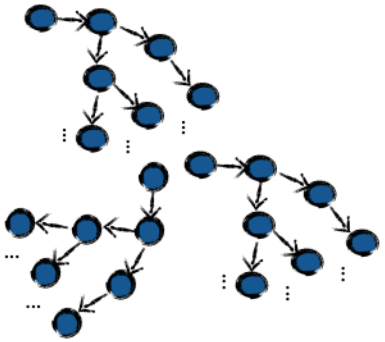
Probability distribution

$$p(\mathbf{v}, \mathbf{h}) \propto \exp\left(-E(\mathbf{v}, \mathbf{h})\right)$$

Weights are trained using contrastive divergence

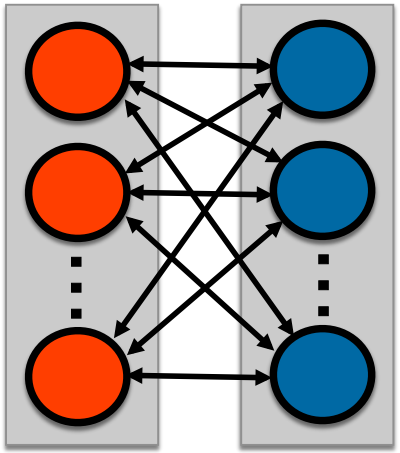
RBDist: Similarity Measure Between MDPs

Step 1: Train an RBM to approximate the source task's dynamics



Sampled trajectories capturing source dynamics

Separate into $\langle s, a, s' \rangle$ i.i.d. tuples and train RBM

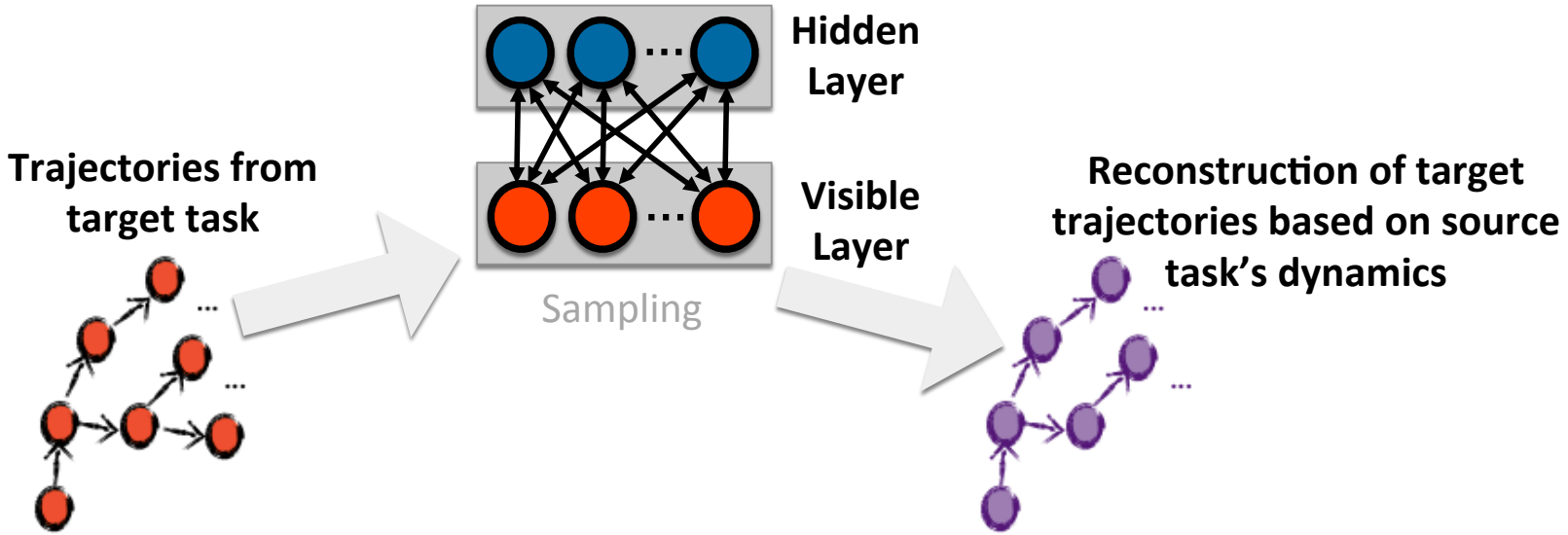


The RBM learns a generative model that captures the source dynamics.

Key Idea: If the dynamics of a source and target domain are similar, the RBM trained on the source task should be able to **reconstruct** trajectories from the target task.

RBDist: Similarity Measure Between MDPs

Step 2: Reconstruct target task trajectories by sampling the trained RBM

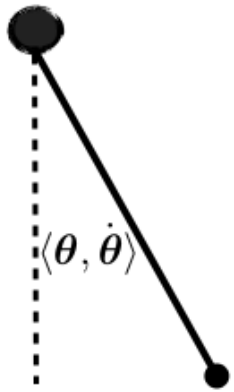


Step 3: Measure reconstruction error of sampled target trajectories as RBDist

$$\text{RBDist} = \frac{1}{n}\sum_{k=1}^{n} e_k \qquad e_k = \text{L}_2\left(\underbrace{\left\langle s_2^{(k)}, a_2^{(k)}, {s'}_2^{(k)} \right\rangle_0}_{\text{original tuple}}, \underbrace{\left\langle s_2^{(k)}, a_2^{(k)}, {s'}_2^{(k)} \right\rangle_1}_{\text{reconstructed tuple}}\right)$$

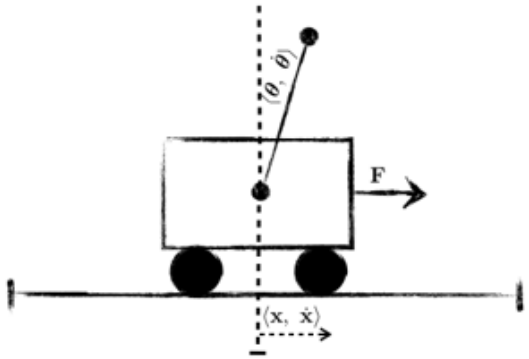
Experiments & Results

Dynamical Systems & Benchmarks



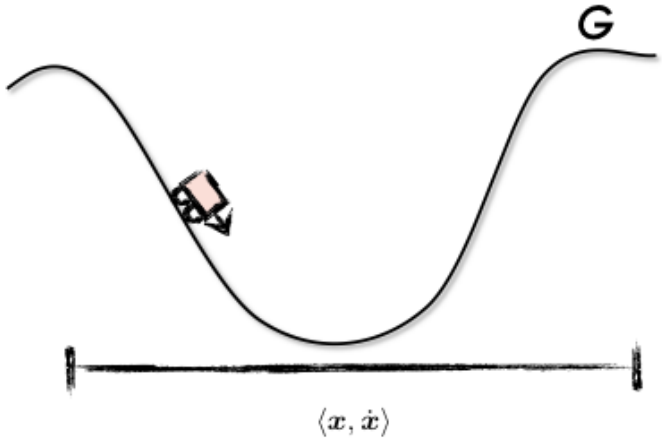
Inverted Pendulum

Swing and balance pole upright by applying torques



Cart Pole

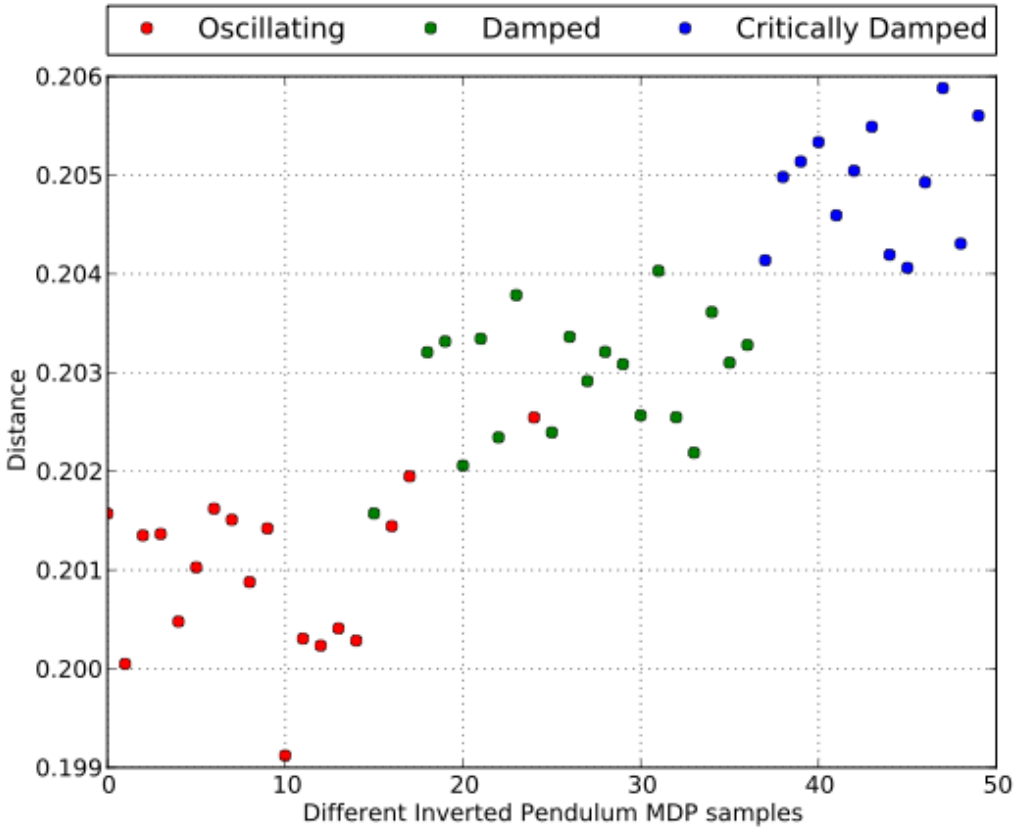
Balance pole upright by applying linear forces



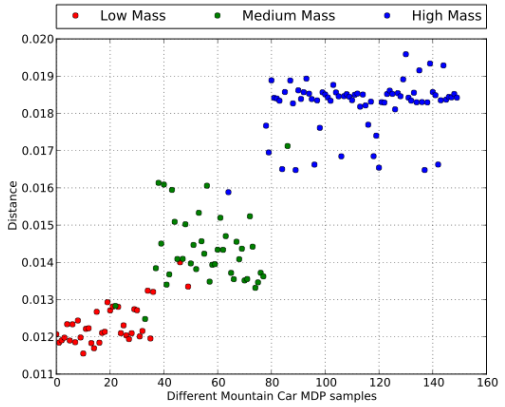
Mountain Car

Control car to reach goal by oscillating around the valley

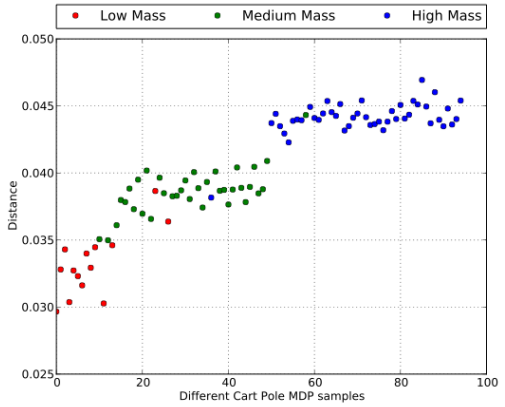
Results: Dynamical Phases



Inverted Pendulum



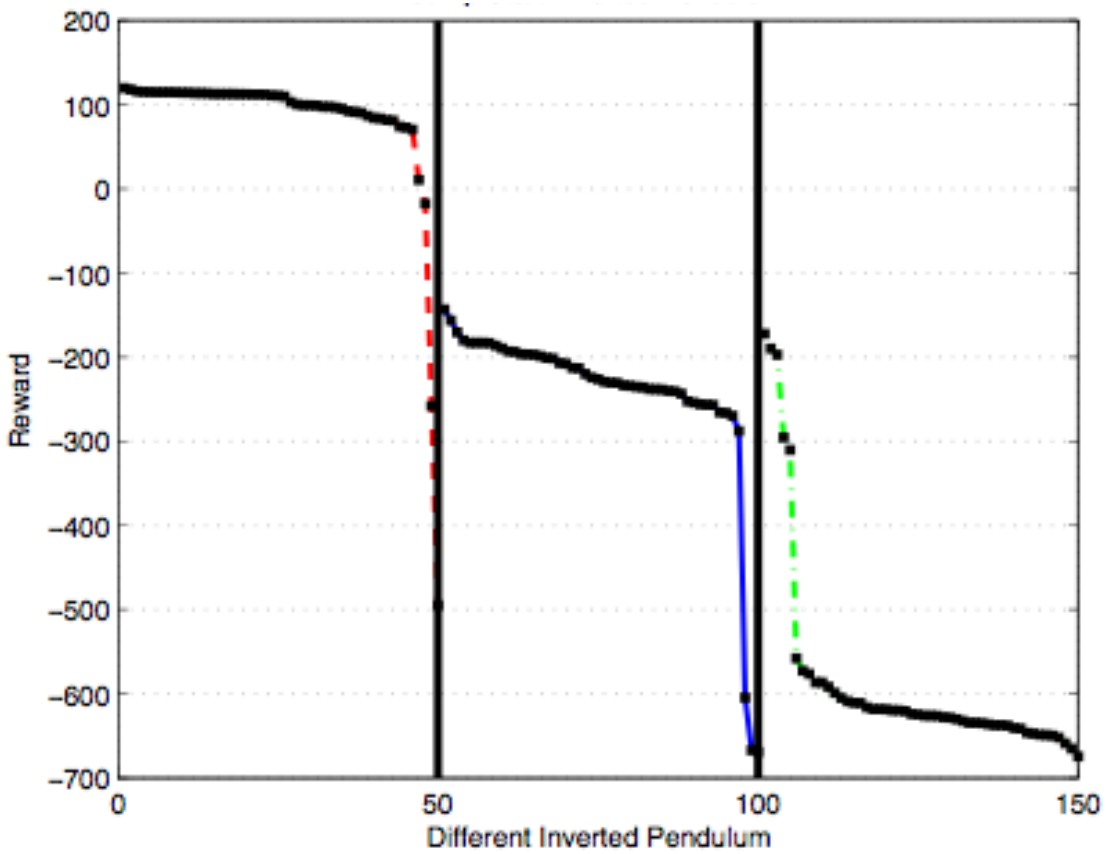
Mountain Car



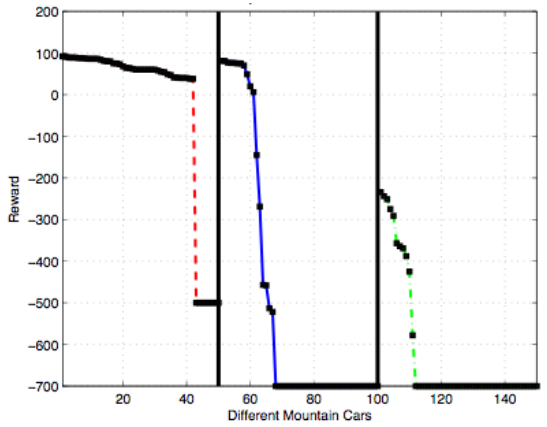
Cart Pole

RBDist can automatically cluster dynamical phases

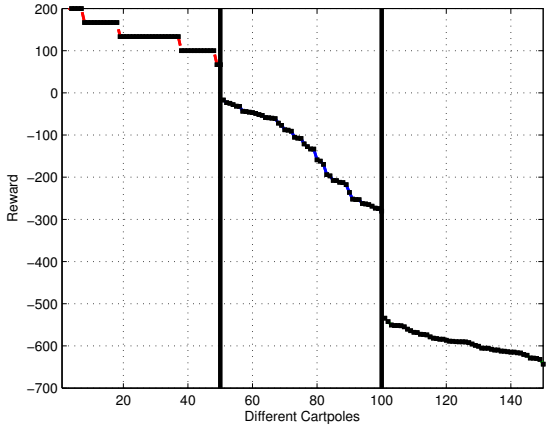
Results: Transfer Performance



Inverted Pendulum



Mountain Car



Cart Pole

RBDist correlates with transfer performance

Thank you!

Please send correspondence to:

Haitham Bou Ammar
haithamb@seas.upenn.edu

Eric Eaton
eeaton@seas.upenn.edu

This work was supported in part by ONR N00014-11-1-0139, AFOSR FA8750-14-1-0069, and NSF IIS-1149917.