Unsupervised Cross-Domain Transfer in Policy Gradient Reinforcement Learning via Manifold Alignment

Haitham Bou Ammar — Univ. Pennsylvania
Eric Eaton — Univ. Pennsylvania
Paul Ruvolo — Olin College
Matthew E. Taylor — Washington State Univ.

Introduction

We developed an autonomous framework that uses unsupervised manifold alignment to learn inter-task mappings and effectively transfer samples between different task domains. Our results demonstrate the success of our approach for transfer between highly dissimilar control tasks (e.g., from cart-poles to quadrotors), and show that transfer quality is positively correlated with manifold alignment quality.

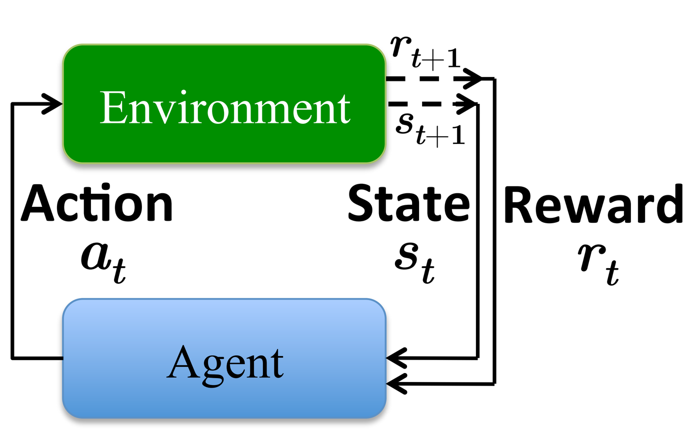
Motivation:

- Transfer learning enables rapid training of a control policy for a new target task by reusing knowledge from other source tasks.
- In the case of multiple task domain, an **inter-task mapping** χ is needed to map knowledge between tasks.
  - χ maps state-action-next-state triplets from the source task to the target task, which can be used for policy initialization.

Background: Reinforcement Learning

Reinforcement Learning (RL) problems are formalized as Markov Decision Processes (MDPs): $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}_0, \mathcal{P}, r \rangle$, where

- $\mathcal{S} \in \mathbb{R}^d$ is the state space
- $\mathcal{A} \in \mathbb{R}^m$ is the action space
- $\mathcal{P}_0$ is the initial state distribution
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0,1]$ is the transition probability function
- $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the reward function.



Goal: Learn an optimal policy $\pi^\star : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the total discounted reward.

Background: Policy Gradient RL

In Policy Gradient (PG) methods, the policy is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ and a vector of state features $\boldsymbol{\Phi}$. The goal is to maximize

$$\mathcal{J}(\boldsymbol{\theta}) = \int_{\mathbb{T}} p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \mathcal{R}(\boldsymbol{\tau}) d\boldsymbol{\tau} \quad \text{, where}$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) = \mathcal{P}_0(\boldsymbol{s}_1) \prod_{t=1}^{H} \mathcal{P}(\boldsymbol{s}_{t+1} | \boldsymbol{s}_t, \boldsymbol{a}_t) \pi(\boldsymbol{a}_t | \boldsymbol{s}_t) \quad \leftarrow \text{Probability of trajectory}$$

$$\mathcal{R}(\boldsymbol{\tau}) = \frac{1}{H} \sum_{t=1}^{H} r(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t, \boldsymbol{s}_t) \quad \leftarrow \text{Reward of trajectory}$$

Problem: PG suffers from high computational and sample complexities.

Unsupervised Manifold Alignment for Learning the Inter-Task Mapping χ_S

Phase I: Learning the inter-task mapping χ_S via unsupervised manifold alignment

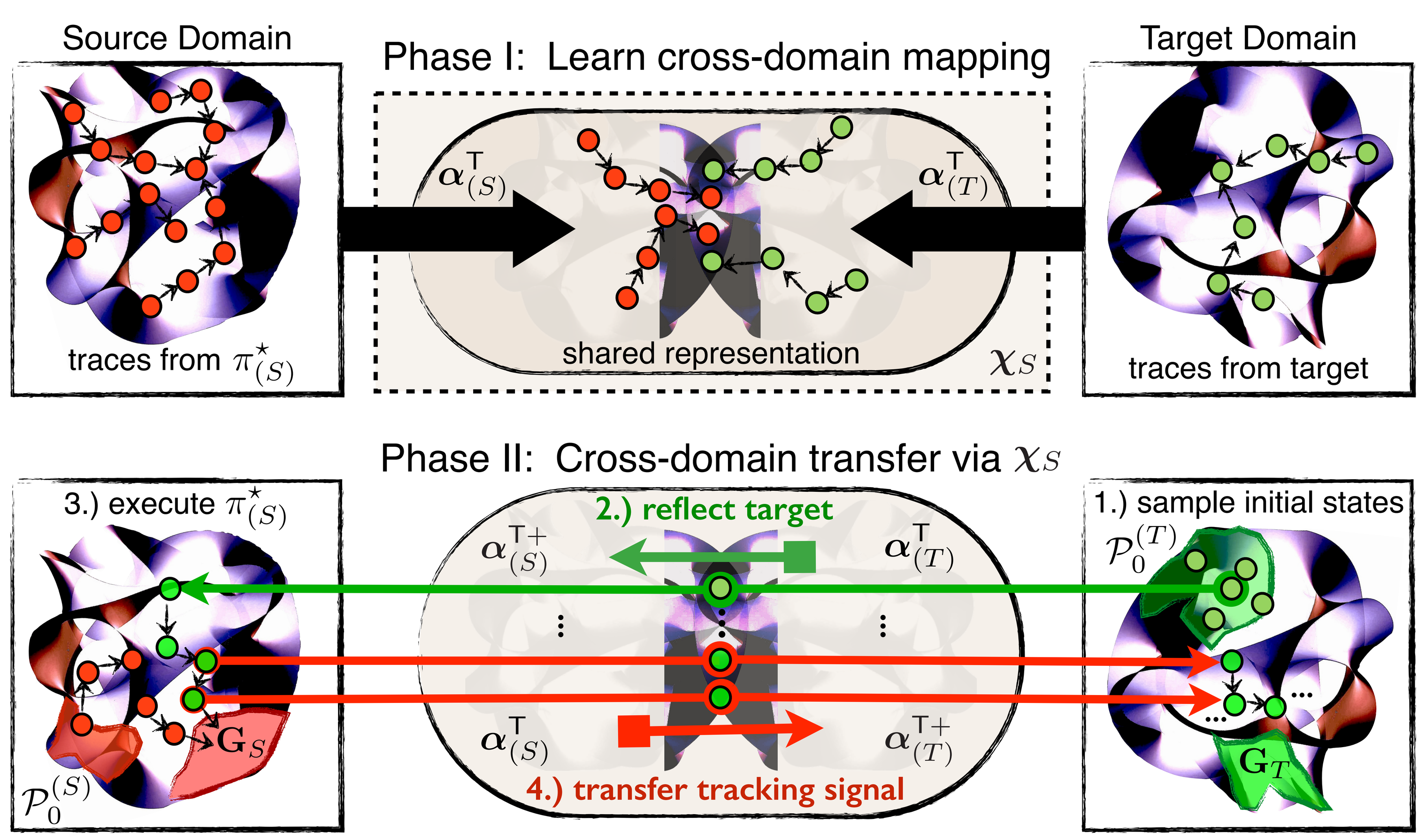
1. Sample *a.)* *optimal* trajectories from the source task using $\pi^\star_{(S)}$ and *b.)* random trajectories from the target task.
2. Flatten all trajectories and construct a k-NN graph to capture the local geometry of the states in both the source and target tasks.
3. Identify a shared representation between the source and target tasks that captures local state transition dynamics by optimizing

$$\mathcal{J}\left(\boldsymbol{\alpha}_{(S)}, \boldsymbol{\alpha}_{(T)}\right) = \underbrace{\mu \sum_{i,j} \left(\boldsymbol{\alpha}_{(S)}^{\mathsf{T}} \boldsymbol{s}_i^{(S)\star} - \boldsymbol{\alpha}_{(T)}^{\mathsf{T}} \boldsymbol{s}_j^{(T)}\right)^2 \boldsymbol{W}_{i,j}}_{\text{cross-task connection}} + \underbrace{0.5 \sum_{i,j} \left(\boldsymbol{\alpha}_{(S)}^{\mathsf{T}} \boldsymbol{s}_i^{(S)\star} - \boldsymbol{\alpha}_{(S)}^{\mathsf{T}} \boldsymbol{s}_j^{(S)}\right)^2 \boldsymbol{W}_{\mathcal{S}^{(S)}}^{i,j}}_{\text{source task geometry}} + \underbrace{0.5 \sum_{i,j} \left(\boldsymbol{\alpha}_{(T)}^{\mathsf{T}} \boldsymbol{s}_i^{(T)\star} - \boldsymbol{\alpha}_{(T)}^{\mathsf{T}} \boldsymbol{s}_j^{(T)}\right)^2 \boldsymbol{W}_{\mathcal{S}^{(T)}}^{i,j}}_{\text{target task geometry}} ,$$

   where the $\boldsymbol{W}$'s are the weighted adjacency matrices, $\boldsymbol{s}$ are the states, the $\boldsymbol{\alpha}$'s are the projections into the shared latent space, and the superscripts or subscripts of (S) and (T) denote whether these variables correspond to the source or target task, respectively.
4. The inter-task mapping $\chi_S = \boldsymbol{\alpha}_{(T)}^{\mathsf{T}+} \boldsymbol{\alpha}_{(S)}^{\mathsf{T}}[\cdot]$.

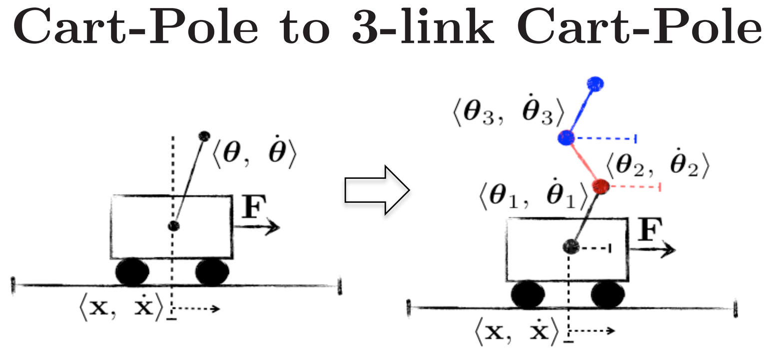
Phase II: Initialize the target task's policy via transfer

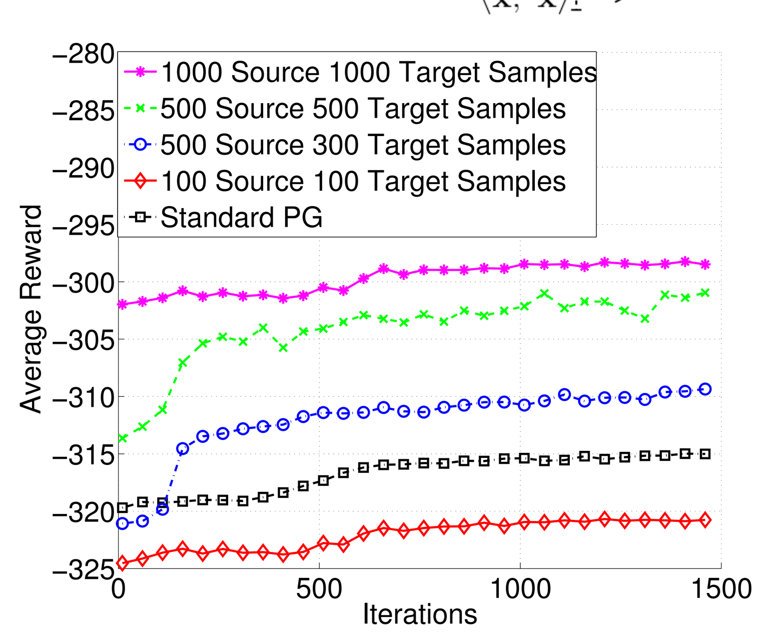
1. Sample initial target states $\boldsymbol{s}_0^{(T)} \sim \mathcal{P}_0^{(T)}$.
2. Project initial target states $\boldsymbol{s}_0^{(T)}$ to the source task via χ_S.
3. Execute $\pi^\star_{(S)}$ from these projected states, yielding optimal trajectories $\tilde{\boldsymbol{\tau}}_{(S)}$.
4. Transfer optimal source trajectories $\tilde{\boldsymbol{\tau}}_{(S)}$ to the target task via χ_S^+, yielding target trajectories $\tilde{\boldsymbol{\tau}}_{(T)}$.
5. Initialize target task policy $\boldsymbol{\pi}_{(T)}$ from $\tilde{\boldsymbol{\tau}}_{(T)}$, yielding $\boldsymbol{\theta}_{(T)}^{(0)}$. Improve $\boldsymbol{\pi}_{(T)}$ using standard policy gradient methods.



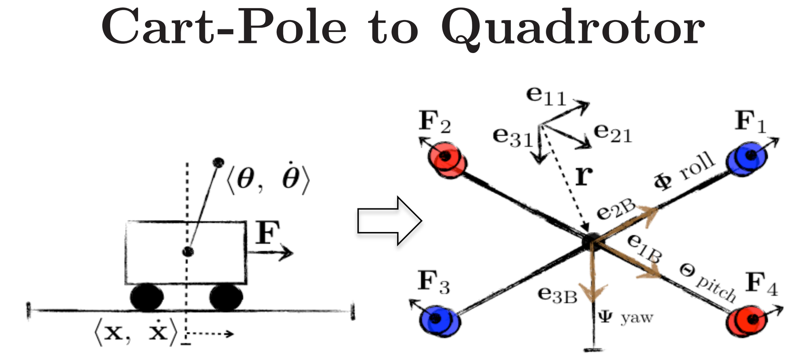
Selected Results of Transfer Between Different Dynamical Systems

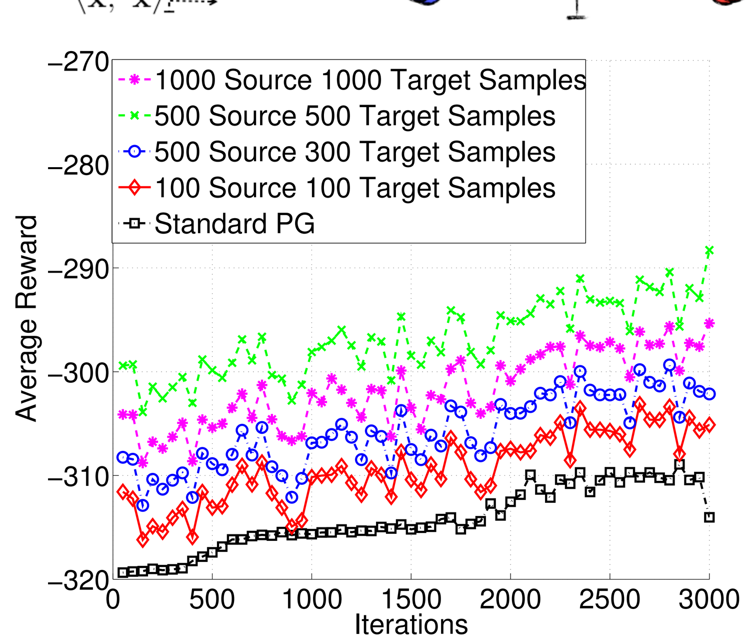
Cart-Pole to 3-link Cart-Pole





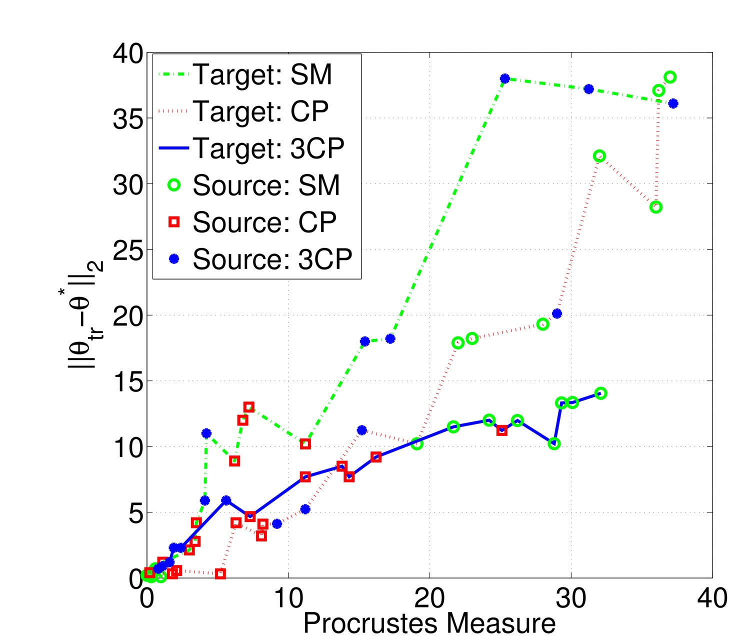
Cart-Pole to Quadrotor





Predicting Success of Cross-Domain Transfer

- Transfer quality ($\|\theta_{tr} - \theta^\star\|_2$) is positive correlated with manifold alignment quality (Procrustes measure).
- Manifold alignment quality may indicate when our approach to cross-domain transfer is likely to succeed.



Unsupervised manifold alignment enables robust cross-domain transfer between highly dissimilar tasks

Acknowledgements: This research was supported in part by ONR grant #N00014-11-1-0139, AFOSR grant #FA8750-14-1-0069, and NSF grant IIS-1149917.