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

Haitham Bou Ammar Eric Eaton Paul Ruvolo Matthew E. Taylor
Univ. Pennsylvania Univ. Pennsylvania Olin College 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 $\chi$ is needed to map knowledge between tasks.
- $\chi$ 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): $(S, A, P_0, P, r)$, where

- $S \in \mathbb{R}^d$ is the state space
- $A \in \mathbb{R}^m$ is the action space
- $P_0$ is the initial state distribution
- $P : S \times A \times S \rightarrow [0, 1]$ is the transition probability function
- $r : S \times A \times S \rightarrow \mathbb{R}$ is the reward function.

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

**Background:**

**Policy Gradient RL**

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

$$\mathcal{J}(\theta) = \int p_\theta(\tau) R(\tau) d\tau,$$

where

$$p_\theta(\tau) = P_0(s_0) \prod_{t=1}^H P(s_{t+1}|s_t, a_t) \pi(\alpha_t|s_t) \quad \text{← Probability of trajectory}$$

$$R(\tau) = \frac{1}{H} \sum_{t=1}^H r(s_{t+1}, a_t, s_t) \quad \text{← Reward of trajectory}$$

**Problem:** PG suffers from high computational and sample complexities.

---

**Unsupervised Manifold Alignment for Learning the Inter-Task Mapping $\chi_S$**

**Phase I:** Learning the inter-task mapping $\chi$ via unsupervised manifold alignment

1. Sample $a_s$) optimal trajectories from the source task using $\pi_s^*$ and $a_t$) 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

$$\mathcal{J}(\alpha_s, \alpha_t) = \mu \sum_{i,j} (\alpha_s^{T(s_i)} \alpha_s^{T(s_j)} - \alpha_t^{T(s_i)} \alpha_t^{T(s_j)})^2 W_{ij} + 0.5 \sum_{i,j} (\alpha_s^{T(s_i)} \alpha_s^{T(s_j)} - \alpha_t^{T(s_i)} \alpha_t^{T(s_j)})^2 W_{S(i,j)} + 0.5 \sum_{i,j} (\alpha_s^{T(s_i)} \alpha_s^{T(s_j)} - \alpha_t^{T(s_i)} \alpha_t^{T(s_j)})^2 W_{S(i,j)} ,$$

where the $W$’s are the weighted adjacency matrices, $s$ are the states, the $\alpha$’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 $\chi_s = \alpha_t^{T(s_i)} \alpha_t^{T(s_j)}$.

**Phase II:** Initialize the target task’s policy via transfer

1. Sample initial target states $s_0^{(T)} \sim \pi_0^{(T)}$.
2. Project initial target states $s_0^{(T)}$ to the source task via $\chi$.
3. Execute $\pi_s^*$ from these projected states, yielding optimal trajectories $\tau_s^{(T)}$.
4. Transfer optimal source trajectories $\tau_s$ to the target task via $\chi_s$, yielding target trajectories $\tau_T^{(T)}$.
5. Initialize target task policy $\pi_T^{(T)}$ from $\tau_T^{(T)}$, yielding $\theta_T^{(0)}$.

Improve $\pi_T^{(T)}$ 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 (1/||$\theta_0 - \theta^*||_1$) 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.

---

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.