Pessimism Can Be Effective: Towards a Framework for Zero-Shot Transfer Reinforcement Learning

Chi Zhang[1], Ziying Jia[2], George K. Atia[1,3], Sihong He[2], Yue Wang[1,3]

[1]Department of Electrical and Computer Engineering, University of Central Florida

[2]Department of Computer Science, University of Texas at Arlington

[3]Department of Computer Science and Engineering, University of Central Florida

Zero-Shot Multi-Domain Transfer Learning

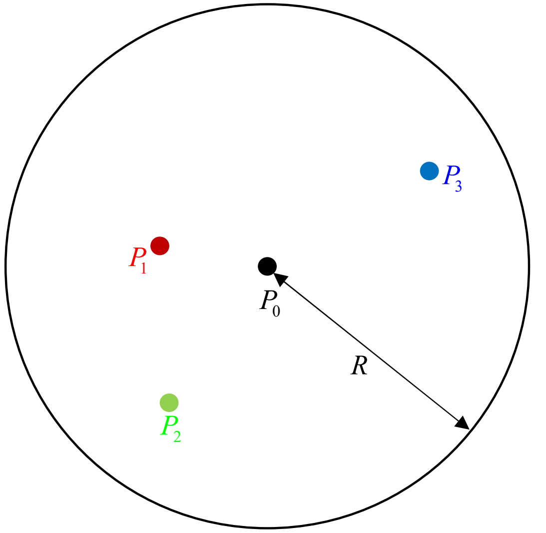
- **Goal**

  Transfer knowledge from diverse and heterogeneous environments without access to target domain

- **Challenges**
  - **Sim-to-Real gap:** discrepancy between target and source domains
  - **Negative transfer:** misleading information contributed by irrelevant source domains
  - **Privacy:** raw data sharing is prohibited

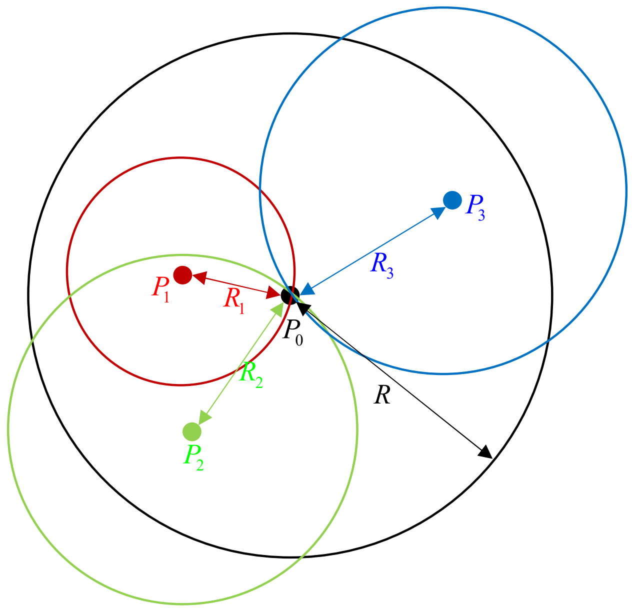
Problem Formulation

- The goal of zero-shot transfer learning is to optimize the performance under P_0, which is not accessible for training
- There are related source domains within an uncertainty set centered at P_0 with radius R



Research Methodology

- **Central Server:** collects and aggregates information from source domains and distributes updates back to local agents
- **Local Agents:** performs robust learning based on the received updates and returns a conservative proxy to the central server
- The local uncertainty set for each agent is constructed to contains P_0



Average Operator-Based Proxy

- **Robust Local Updates:** $Q_k(s,a) \leftarrow (1-\lambda)Q_k(s,a) + \lambda(r(s,a) + \gamma\sigma_{(\mathcal{P}_k)_s^a}(V_k))$

- **Global Aggregation:** $\bar{Q}(s,a) \leftarrow \frac{1}{K}\sum_{k=1}^{K} Q_k(s,a)$

Average Operator-Based Proxy

- The proxy yields a lower bound on the performance of the policy under the target environment
- The proxy is less conservative than proximal robust domain randomization (a direct extension of domain randomization)
- By reducing the aggregation frequency, the convergence rate of our method enjoys a partial linear speedup w.r.t. the number of the agents

$$\left\| \mathbb{E}\left[Q_{\mathrm{AO}} - \frac{\sum_{k=1}^{K} Q_k}{K} \right] \right\| \leq \tilde{O}\left(\frac{1}{T\boxed{K}} + \frac{(E-1)\Gamma}{T} \right)$$

Minimal Pessimism Principle

- **Robust Local Updates:** $Q_k(s,a) \leftarrow (1-\lambda)Q_k(s,a) + \lambda(r(s,a) + \gamma\sigma_{(\mathcal{P}_k)_s^a}(V_k))$

- **Global Aggregation:** $\hat{Q}(s,a) \leftarrow \max_{k\in\mathcal{K}} Q_k(s,a)$

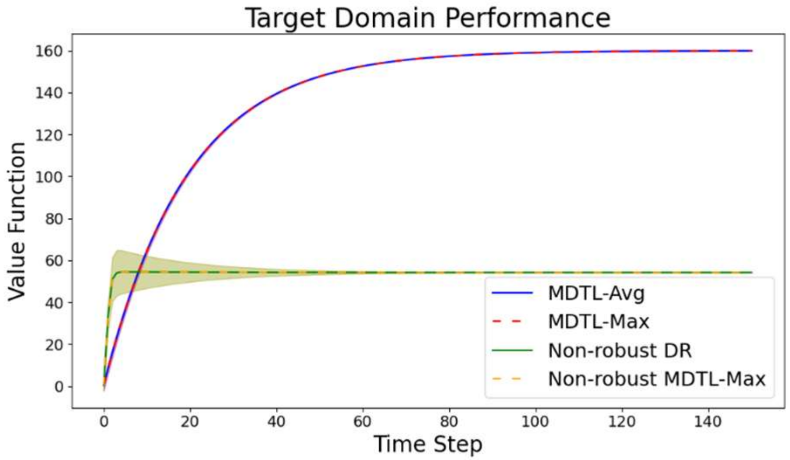
Minimal Pessimism Principle

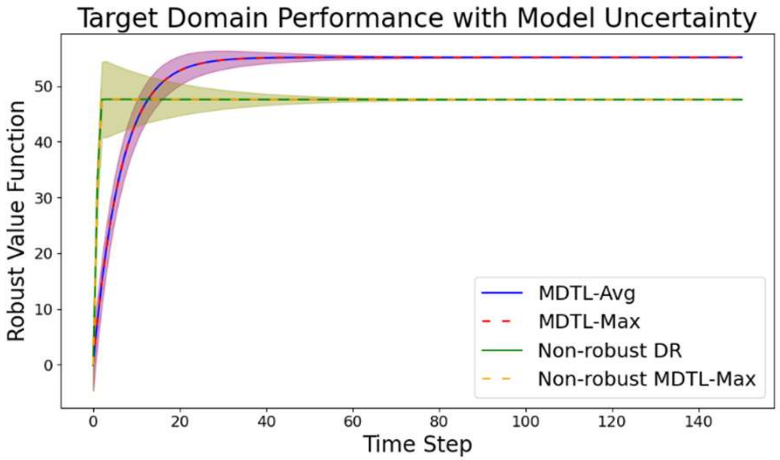
- The proxy yields a lower bound on the performance of the policy under the target environment
- The proxy is less conservative than the robust value function of any source domain and the average operator-based proxy, and hence avoids negative transfer
- By reducing the aggregation frequency, the convergence rate of our method enjoys a partial linear speedup w.r.t. the number of the agents

$$\left\| \mathbb{E}\left[Q_{\mathrm{MP}} - \max_{k \in \mathcal{K}} Q_k \right] \right\| \leq \tilde{\mathcal{O}}\left(\frac{1}{T\boxed{K}} + \frac{(E-1)\Gamma}{T} \right)$$

Experimental Results

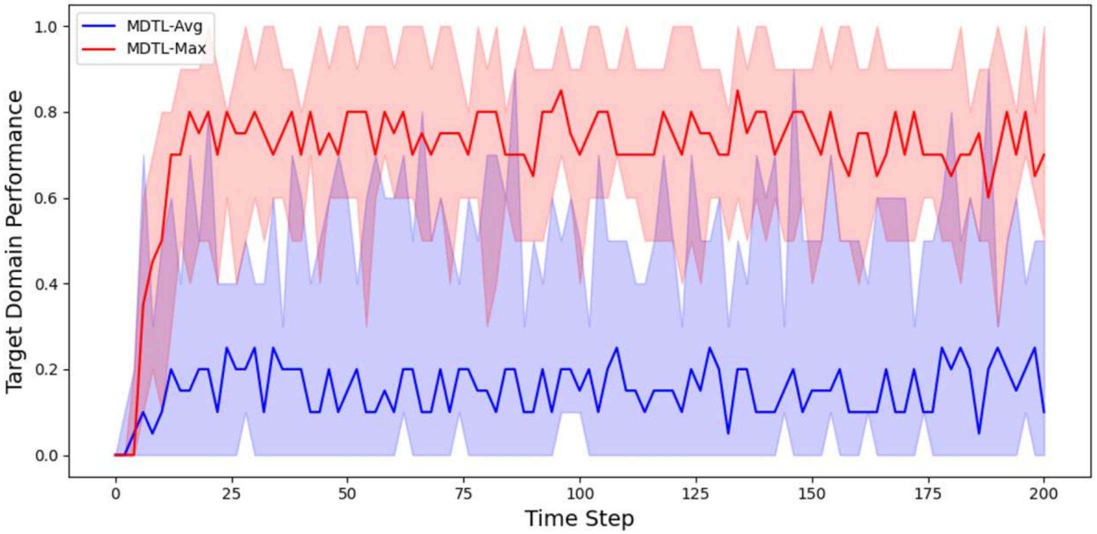
- Recycling Robot





Experimental Results

- Effect of Negative Transfer



Thanks for Watching!