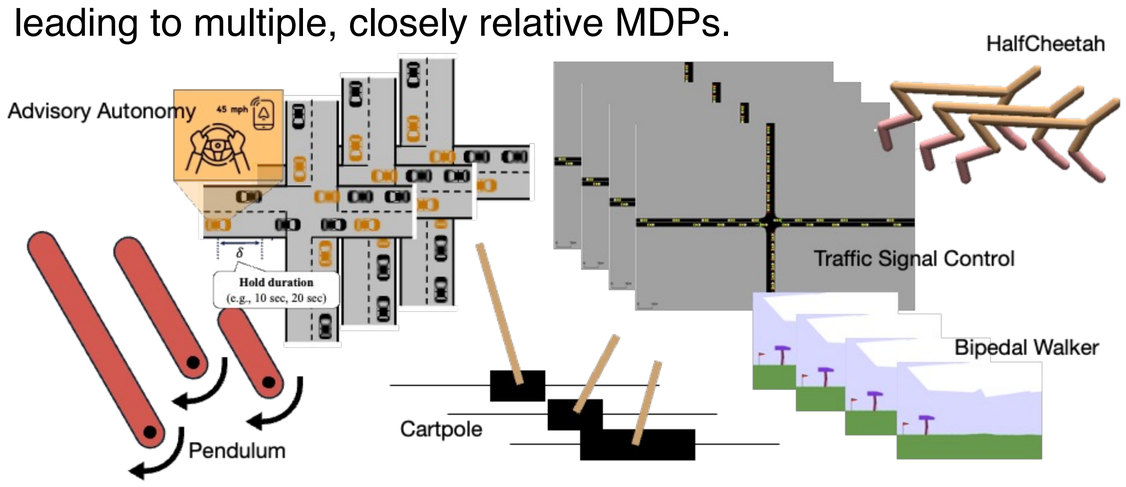
Structure Detection for Contextual Reinforcement Learning

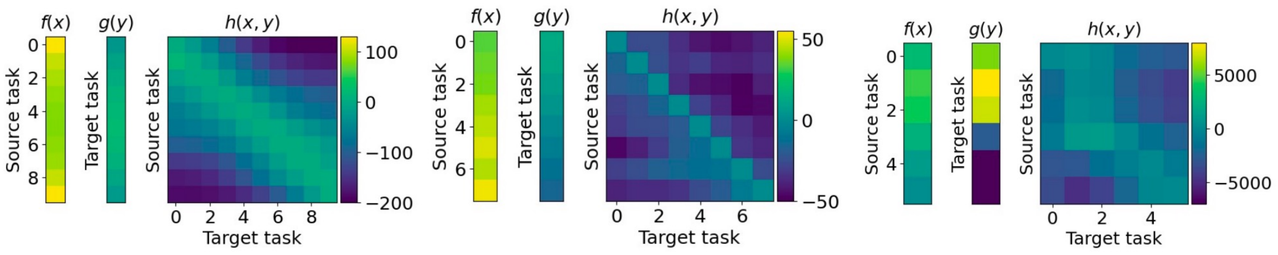
Tianyue Zhou*, Jung-Hoon Cho*, Cathy Wu
Massachusetts Institute of Technology
*These authors contributed equally

Introduction

- **Goal:** Solve **Contextual Markov Decision Process (CMDP)** which extends MDPs with additional contextual information, leading to multiple, closely relative MDPs.



- **Multi-policy training** is a promising paradigm. Model-Based Transfer Learning (**MBTL**) is a multi-policy approach that **strategically selects training tasks** and then **zero-shot transfers** the learned policies to target tasks
- **Limitation & challenge**: Only considers **1D** contexts, while multi-dim demands **significantly more data**
- **Motivation:** Detect structures in CMDP. Exploit the structure to **curb unnecessary exploration.**
- **Heatmaps of decomposed generalization performance:**



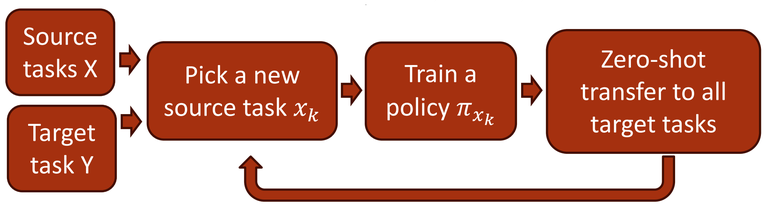
(a) CartPole (b) BipedalWalker (c) CyclesGym

Problem Formulation & Previous Work

- Contexts of target tasks: $Y = \{y_1, y_2, \ldots, y_N\}$
- Each MDP: $\mathcal{M}_y = (S, A, P_y, R_y, \rho_y)$
- Goal:

$$x_k = \arg\max_{x \in X \setminus \{x_{1:k-1}\}} \mathbb{E}_{y \sim \mathcal{U}(Y)} [\max_{x' \in x_{1:k-1} \cup \{x\}} J(\pi_{x'}, y)]$$

- Where: $x_{1:K} = x_1, \ldots, x_K$



- Gaussian Process MBTL (**GP-MBTL**, NeurIPS 2024) achieves up to **40x** more **sample efficient** than independent & multi-task baselines:

$$J(\pi_x, y) \approx J(\pi_x, x) - |x - y|$$

Modeled by Gaussian process **Linear generalization gap (1D)**

Experiments

Benchmark (CMDP)	Independent	Multi-task	Random	GP-MBTL	M-MBTL (Ours)	M/GP-MBTL (Ours)	Myopic Oracle
CartPole ($K = 12$)	0.9346 ± 0.0003	**0.9967** ± 0.0024	0.9861 ± 0.0017	0.9919 ± 0.0013	0.9896 ± 0.0016	0.9898 ± 0.0016	0.9998 ± 0.0000
BipedalWalker ($K = 12$)	0.7794 ± 0.0011	0.5680 ± 0.0919	0.8051 ± 0.0045	0.8073 ± 0.0044	**0.8315** ± 0.0029	**0.8261** ± 0.0030	0.8629 ± 0.0011
IntersectionZoo ($K = 50$)	0.2045 ± 0.0008	0.3788 ± 0.1059	0.5288 ± 0.0108	**0.5840** ± 0.0092	0.4878 ± 0.0064	**0.5682** ± 0.0069	0.6305 ± 0.0082
CyclesGym ($K = 50$)	0.2133 ± 0.0002	0.2081 ± 0.0002	0.2198 ± 0.0001	0.2193 ± 0.0001	**0.2205** ± 0.0001	0.2201 ± 0.0001	0.2214 ± 0.0001
Aggregated Performance	-2.9063 ± 0.1470	-3.1120 ± 1.8408	0.0000 ± 0.0467	0.1681 ± 0.0539	0.1822 ± 0.0503	**0.2930** ± 0.0403	1.0000 ± 0.0236

12.49% aggregated improvement

- M/GP-MBTL **matches the stronger of M- and GP-MBTL** and **successfully detects to the right algorithm** !

Generalization Structure Decomposition

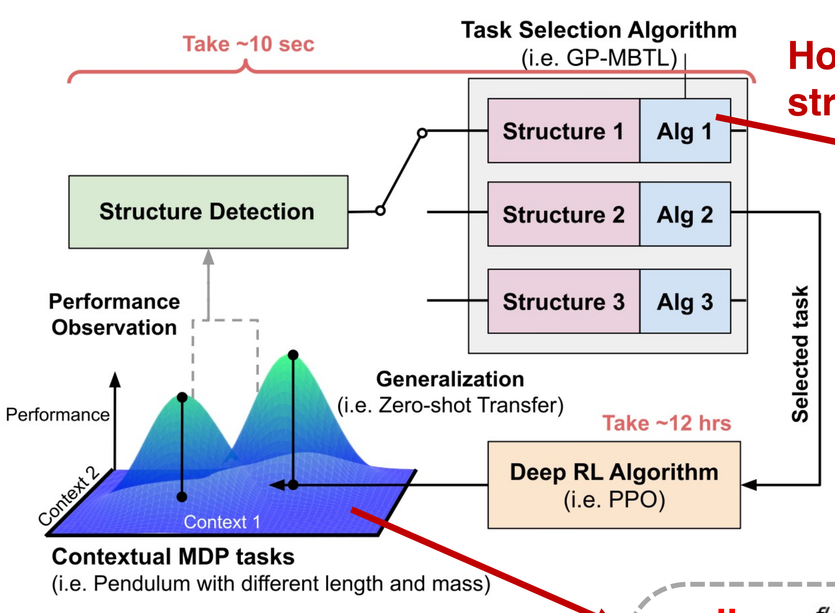
- Inspired by the Sobol-Hoeffding Decomposition:

$$C := \mathbb{E}_{x \in X, y \in Y}[J(\pi_x, y)] \qquad g(y) := \mathbb{E}_{x \in X}[J(\pi_x, y)] - C$$
$$f(x) := J(\pi_x, x) - g(x) - C \qquad h(x, y) := J(\pi_x, y) - f(x) - g(y) - C$$

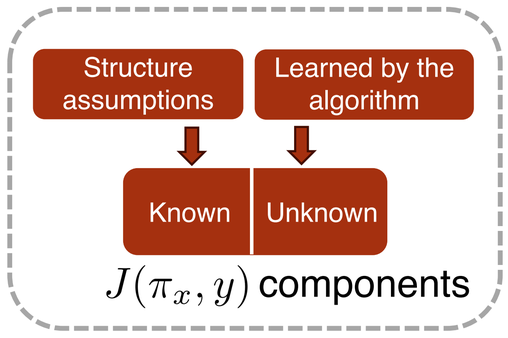
- We can get:

$$J(\pi_x, y) = \underbrace{f(x)}_{\text{policy quality}} + \underbrace{g(y)}_{\text{task difficulty}} + \underbrace{h(x, y)}_{\text{task dissimilarity}} + \underbrace{C}_{\text{constant}}$$

Structure Detection Model-Based Transfer Learning



How to exploit a certain structure?



Mountain Structure

Structure assumptions:

(1) Constant f(x):

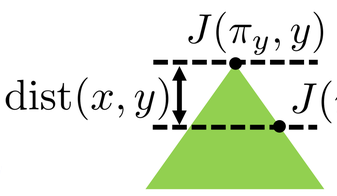
$$f(x) = C$$

(2) Distance Metric h(x,y):

$$h(x, y) = -\text{dist}(x, y)$$

(1) + (2) = Mountain:

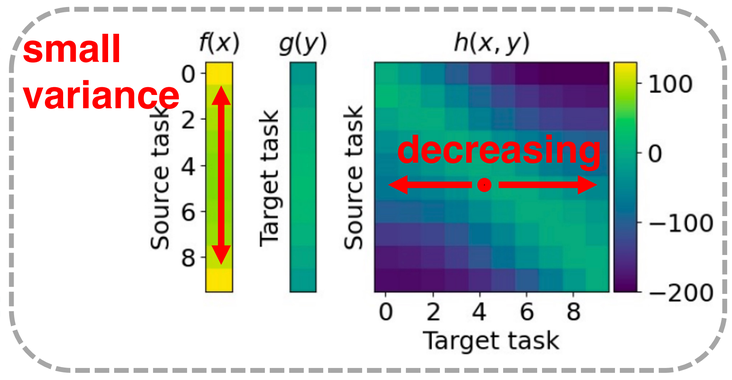
$$J(\pi_x, y) = J(\pi_y, y) - \text{dist}(x, y)$$



Like a mountain!

Unknown: g(y):

$$g(y) \approx \mathbb{E}_{x \in x_{1:k}}[J(\pi_x, y)] - C$$



Structure Detection

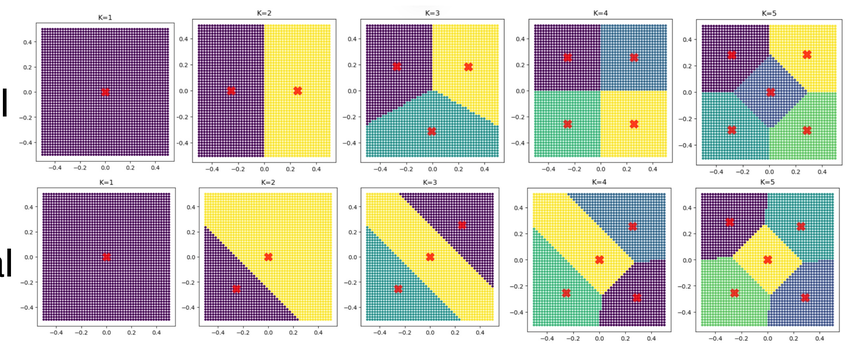
- Criterion 1: f(x) has a relatively small variance
- Criterion 2: The majority of slope signs in h(x,y) are decreasing

Mountain Model-Based Transfer Learning (M-MBTL)

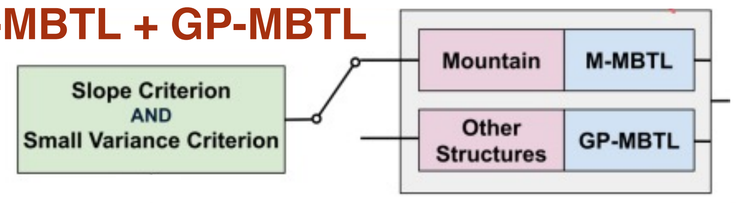
- With Mountain structure, the problem reduces to the **sequential version of clustering**
- M-MBTL extends K-Means to sequential setting, where each centroid represents a training task
- Example in 2D context space:

Non-sequential

Sequential



M/GP-MBTL = M-MBTL + GP-MBTL



Massachusetts Institute of Technology

Structure Detection MBTL (SD-MBTL) dynamically identifies the underlying generalization structure of CMDP and selects appropriate algorithm.
(12.49% aggregated improvement than strongest prior method)