Lecture 11b: Multi-Task RL
CS885 Reinforcement Learning

2022-10-24

Complementary readings:
Vithayathil Varghese, N., & Mahmoud, Q. H. (2020). A survey of multi-task deep reinforcement learning. Electronics, 9(9), 1363.

Pascal Poupart
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO

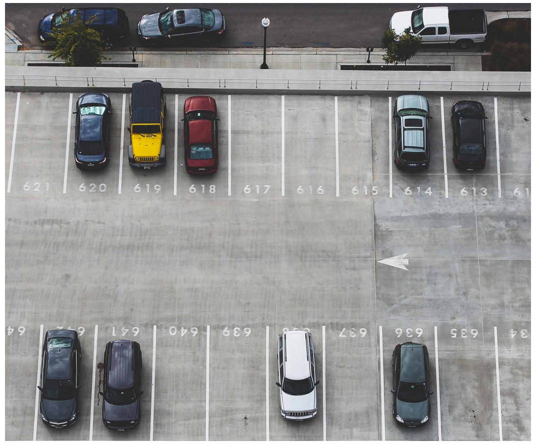
Outline

Transfer across multiple tasks

- Domain adaptation
- Fine tuning
- Randomization
- Contextualized RL

Motivation

- Autonomous driving: Parking in different parking lots

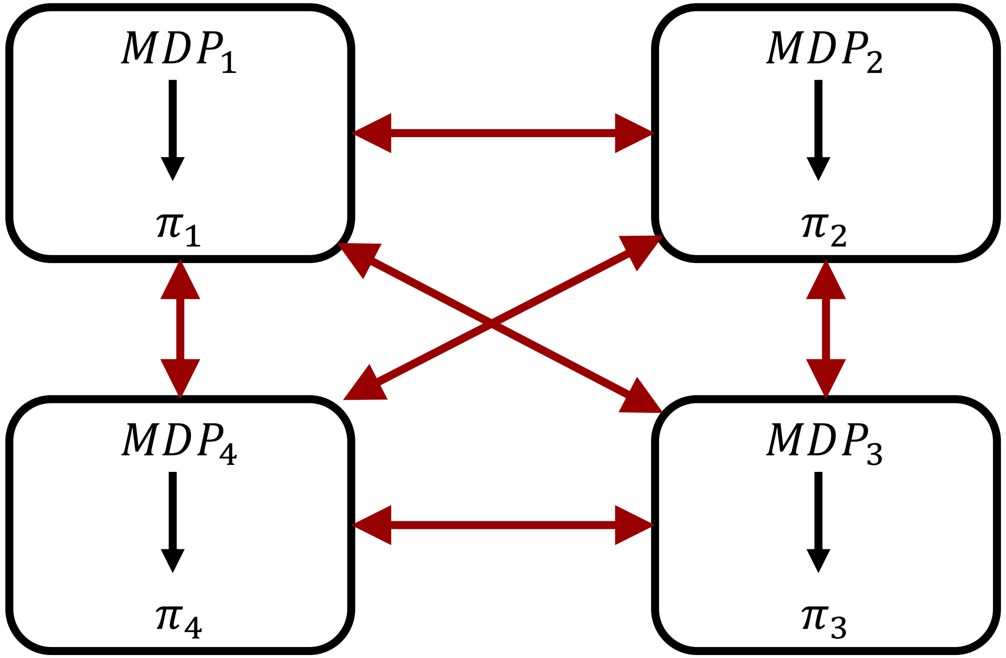


- Conversational agents: answer customers of different clients

UNIVERSITY OF WATERLOO

Multi-Task RL

- Transfer what is learned across tasks (task = MDP)



Transfer Learning

RL task #1:
- States: S_1
- Actions: A_1
- Transitions T_1: $P_1(s'|s,a)$
- Rewards R_1: $P_1(r|s,a)$

Solution:
- Q-function: $Q_1(s,a)$
- Policy: $\pi_1(a|s)$
- Model: $\tilde{T}_1, \tilde{R}_1$

RL task #2:
- States: S_2
- Actions: A_2
- Transitions T_2: $P_2(s'|s,a)$
- Rewards R_2: $P_2(r|s,a)$

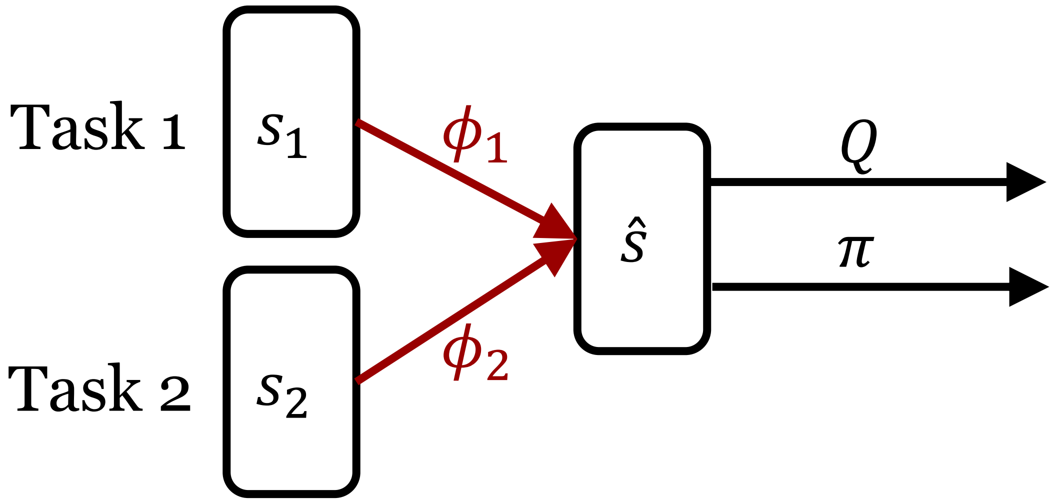
Solution:
- Q-function: $Q_2(s,a)$
- Policy: $\pi_2(a|s)$
- Model: $\tilde{T}_2, \tilde{R}_2$

Techniques for RL Transfer Learning

- Domain adaptation
- Fine tuning
- Randomization
- Contextualized RL

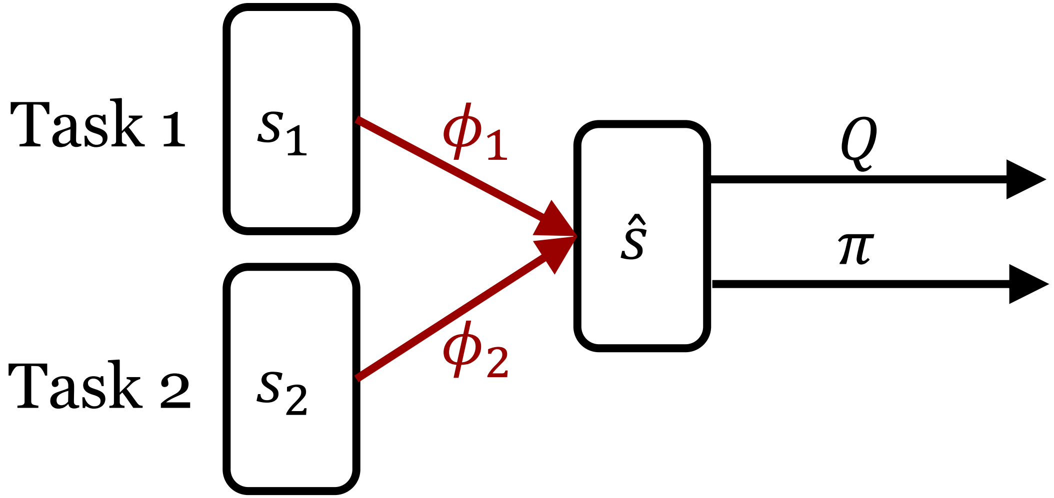
Domain Adaptation

- Learn mappings $\phi_i: s_i \rightarrow \hat{s}$ that find invariant state features $\hat{s}$ that are common across several tasks
  - E.g. Robotics/autonomous driving with different sensors
  - Assumptions: sensors provide same information in a different form; transition and reward models identical



Domain Adaptation

- Actor π and critique Q are shared across tasks
- Simply prepend π and Q with feature mapping ϕ_i of corresponding task when learning with your favorite actor-critic algorithm



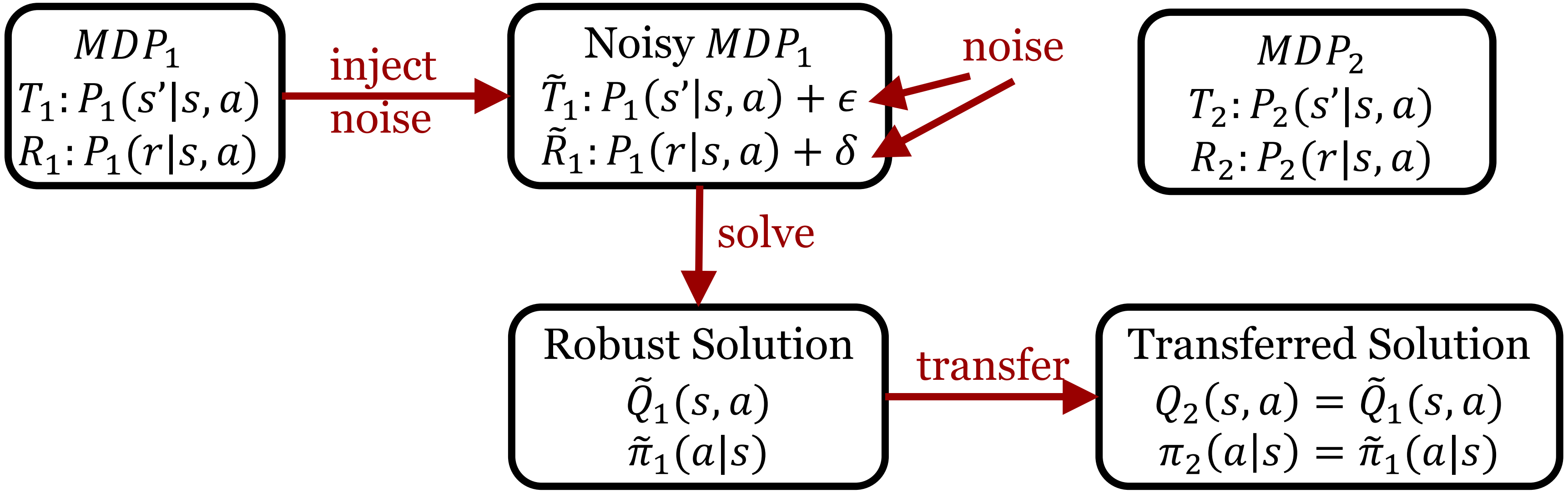
Fine-Tuning

- Pre-train on one task, fine tune on new task
  - Assumption: underlying MDPs are similar

- E.g. conversational agents that answer questions by customers of different clients
  - Task 1 (client 1): Learn Q_1, π_1 with any algorithm
  - Task 2 (client 2): Initialize $Q_2 \leftarrow Q_1$, $\pi_2 \leftarrow \pi_1$ and then continue training Q_2, π_2 with any algorithm
  - Benefit: faster training for Task 2

Task Randomization

- Modify source task by injecting noise/variations to learn robust Q, π for future tasks
  - E.g. sim2real problem: learn policy in simulation to be deployed in real world



Policy Randomization

- Find a maximum entropy policy
  that is likely to generalize to slightly different domains
  - E.g. Soft Q-Learning, Soft Actor-Critic

Contextualized RL

- Augment state s with features g encoding the task
  - Frequent scenario: goal conditioned RL (g refers to goal)

- Examples
  - Robotics/autonomous driving: destination determines reward function of task
  - Automated trading: risk preferences determine reward function of task

UNIVERSITY OF WATERLOO

Contextualized RL

- Augment state s with features g encoding the task
  - Frequent scenario: goal conditioned RL (g refers to goal)

Contextualized RL:

- States: $s \in S$
- Task features: $g \in G$
- Actions: $a \in A$
- Transitions T: $P(s'|s,g,a)$
- Rewards R: $P(r|s,g,a)$

Q-function: $Q(s,g,a)$
Policy: $\pi(a|s,g)$

- Train your favorite RL algorithm
  - Neural net actor-critic will generalize to new g's

Demo: Multi-task RL in Robotics

Gupta, Yu, Zhao, Kumar, Rovinsky, Xu, Devlin, Levine (2021), **Reset-Free Reinforcement Learning via Multi-Task Learning:** Learning Dexterous Manipulation Behaviors without Human Intervention ICRA.

UNIVERSITY OF WATERLOO