Last Time

Imitation Learning

IRL

match π^*

IRL
Given: $S, A, T, \{\tau\}$
Find: R

RL
S, A, T, R
π^*

This Time

Peer Review

Transfer and Meta-Learning

Transfer Learning: Use experience from one set of tasks for faster learning and better performance on a new task.

In RL task = MDP

"source" domain → "target" domain

"shot" = attempt in "target domain"

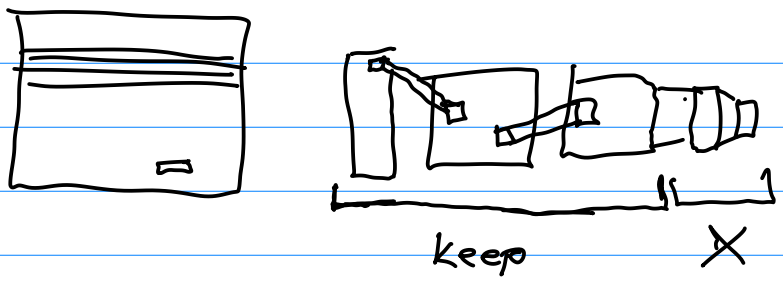
"0-shot" = run policy in target domain

"1-shot" = try task once

"few-shot"

How is prior knowledge stored?

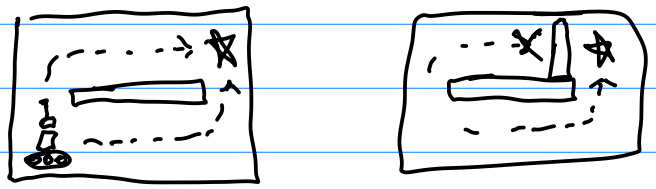
- Features ←
- Policy (Bad)
- Q-function
- Model (Physics)



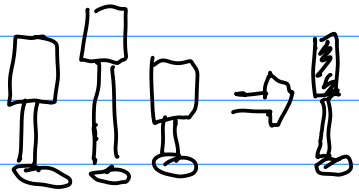
Approaches

Pre-training + Finetuning

Pretrain with robustness and diversity



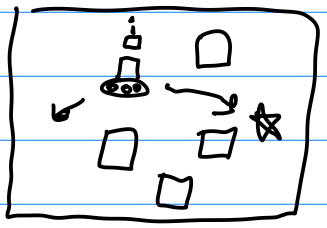
- Soft-Q-learning
- EPOpt



- CAD2RL

Multi-Task RL

Make environment part of the state



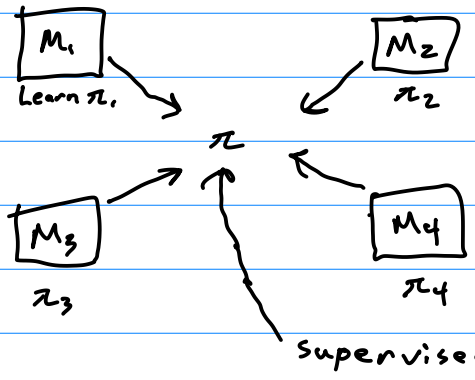
$s = (r_x, r_y, g_x, g_y, o_{1_1}, o_{1_2} \ldots)$

Problems

- Gradient Interference
- Winner-take-all

In practice, often negative transfer

Actor - Mimic



Successor Features - Reason about multiple reward functions at once

Problems all have same S, A, T, γ

different R

Recall $V^\pi = (I - \gamma T^\pi)^{-1} R^\pi$ $|S|$ vector

$Q^\pi = (I - \gamma \bar{T}^\pi)^{-1} \bar{R}$ $|S||A|$ vectors

Let $R(s,a) = \theta^T \phi(s,a)$ ← features

$\bar{\phi}$ feature matrix $|S||A|$ columns, N rows

$\bar{\phi}_{i,k} = \phi_i(s,a)$ ← $k = s_a + a$

→ Successor Feature: $\psi^\pi(s,a) = E\left[\sum_t \gamma^t \phi(s,a)\right]$

$Q^\pi(s,a) = \theta^T \psi^\pi(s,a)$

$\psi^\pi(s,a) = \phi(s,a) + \gamma E_{\substack{s' \sim T^\pi(s,a) \\ a' = \pi(s')}}\left[\psi(s',a')\right]$

$\bar{\psi}^\pi = (I - \gamma \bar{T}^\pi)^{-1} \bar{\phi}$

$$R'(s,a) = \theta'^{T} \phi(s,a)$$

$$Q'^{\pi}(s,a) = \theta'^{T} \psi^{\pi}(s,a)$$

Meta Learning : Learning to Learn in Target domain

RL

$$\theta^* = \arg\max_{\theta} E_{\pi_\theta}[R(\tau)]$$

$$= f_{RL}(M)$$

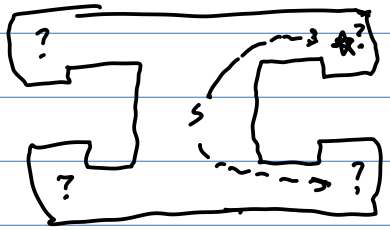
Meta RL

$$\theta^* = \arg\max_{\theta} \sum_i E_{\pi_{\phi_i}}[R(\tau)]$$

where $\phi_i = f_\theta(M_i)$

Approach 1: This is a POMDP

RL^2



Approach 2: Gradient Based Meta RL MAML

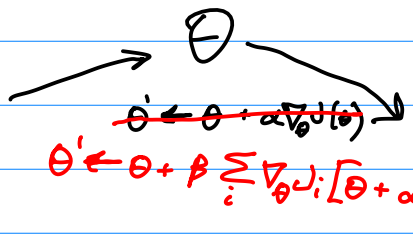
RL (policy gradient)

$$\theta^{k+1} \leftarrow \theta^k + \alpha \nabla_{\theta^k} J(\theta^k)$$

G.B. Meta RL

$$f_\theta(M_i) = \theta + \alpha \nabla_\theta J_i(\theta)$$

Meta



Env
Env
Env

One shot learning