Last time: Exploration

Best: RND
Go-Explore

This time: Transfer, Meta; Grand Accomplishments RL

Transfer Learning : Using experience from one set of tasks for a new task
↖ MDP

Can RL use prior knowledge?

Where to store?

- Q-function
- Policy
- Model
- Features / Hidden States
  - ↳ Aside: Representation Bottleneck

Jargon

"source" → "target"

"shot" = attempt in target domain

"1 shot"
"0 shot"
"few shot"

Forward

MultiTask

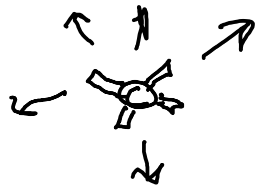
Meta Learning

→ Forward : Fine Tuning

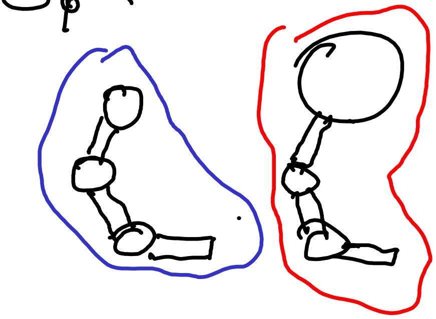
$R(s,a) + H[\pi_\theta]$

Key : Lots of randomness

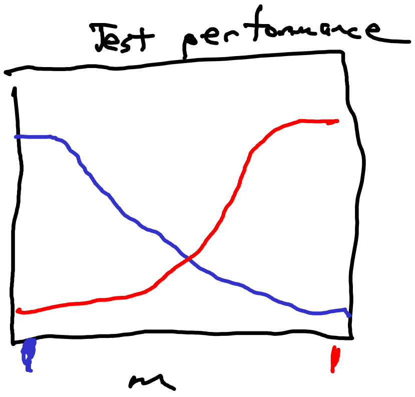
Example

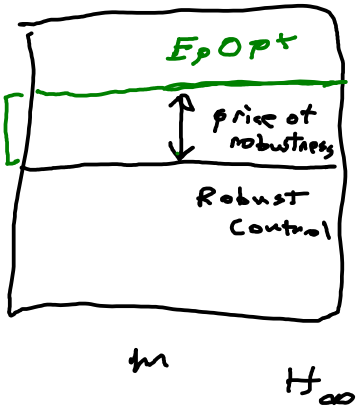


EpOpt



$E[R]$





CAD2RL

→ Multi-Task

Contextual Policy

$\pi_\theta(a|s)$

$\pi_\theta(a|s,\omega)$

where to place an object

direction to run in

$\tilde{S} = S \times \Omega$

Context space

Modular

Meta RL : Learning to Learn

RL

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

$$\theta^* = f_{RL}(M)$$

MDP

RL algorithm

Meta RL

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{n} E_{\pi_{\phi_i}}[R_i(\tau)]$$

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

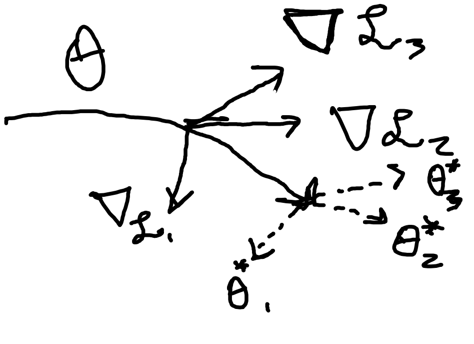
1. sample task i, collect D_i
2. adapt policy $\phi_i = f(\theta, D_i)$
3. Collect D_i' with π_{ϕ_i}
4. Update θ with $\mathcal{L}(D_i', \phi_i)$

3 Solutions

1. Recurrence — RNN

   Example RL^2

2. Optimization — MAML



3. ML as a POMDP

hidden state is task

$\tilde{S} = S \times \mathcal{M}$ ← set of MDPs

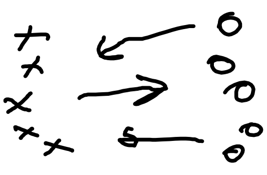
$o_t = (s_t, r_t)$

Learn Filter

Most Celebrated RL Accomplishment

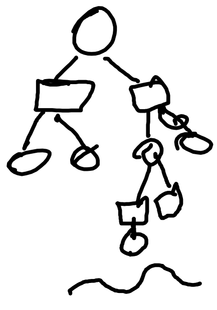
Name	Domain	Online	Deep Learning	Key	Training Time
AlphaGo	Go	MCTS	Policy NN Value NN	Trained on human games, self-play	3 weeks
AlphaZero	Go, chess shogi	MCTS	"	self play	24 hours ←
Open AI 5	Dota	—	PPO with LSTM + expert-specified dense reward	League Play	weeks
Alpha Star	Starcraft II	—	Actor-Critic LSTM TD(λ) returns	League Play	40 days
Deep Stack	Poker.	Heuristic search	Counterfactual Value Network	Incomplete Info	2 days on 1 GPU ←
FTW Agent	Quake CTF	—	Timescale Hierarchical Actor-Critic RNN learned Dense rew.	League Play	500K games × 10 min = 9.5 years
MuZero	Board games Atari	MCTS	Policy Value Model	Self Play	12 hours ←

Dota



Alpha Go 2015

$|A| \approx 19 \times 19 = 361$



Policy Network: suggest a few good action

Value Network: evaluates state nodes at leaves