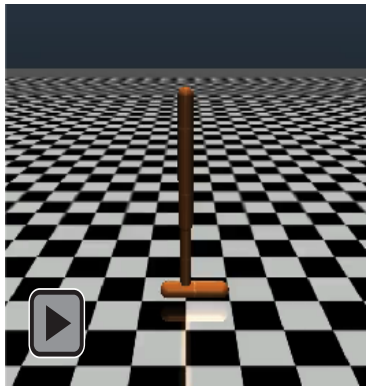
Guided Meta-Policy Search

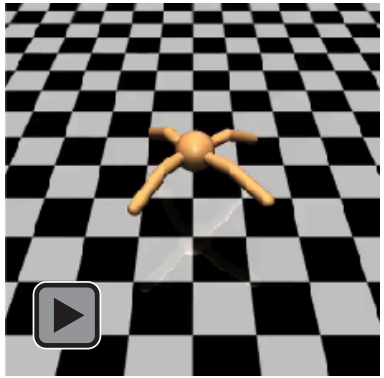
Russell Mendonca, Abhishek Gupta, Rosen Kralev, Pieter Abbeel,
Sergey Levine, Chelsea Finn

High Sample Complexity of RL

Hopper - v1

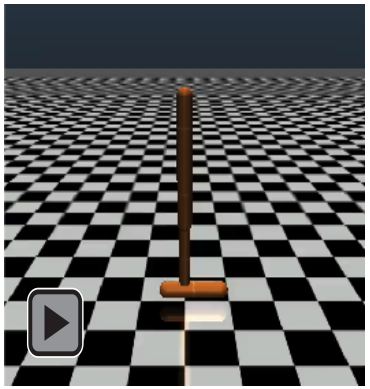


Ant - v1

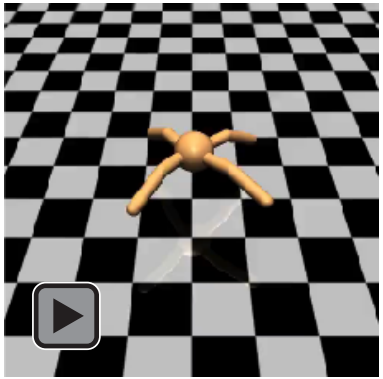


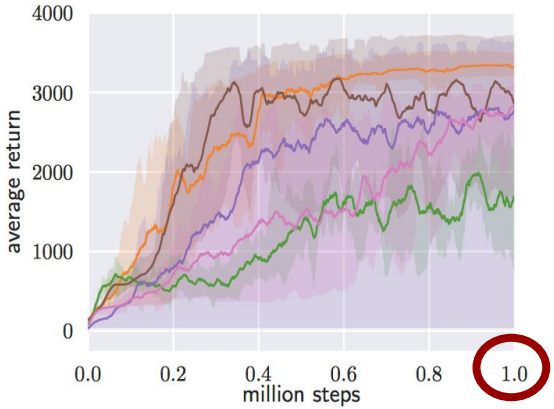
High Sample Complexity of RL

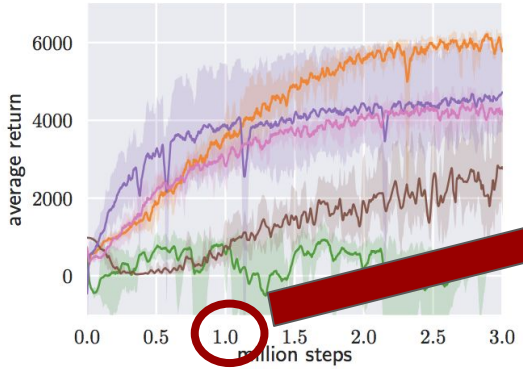
Hopper - v1



Ant - v1

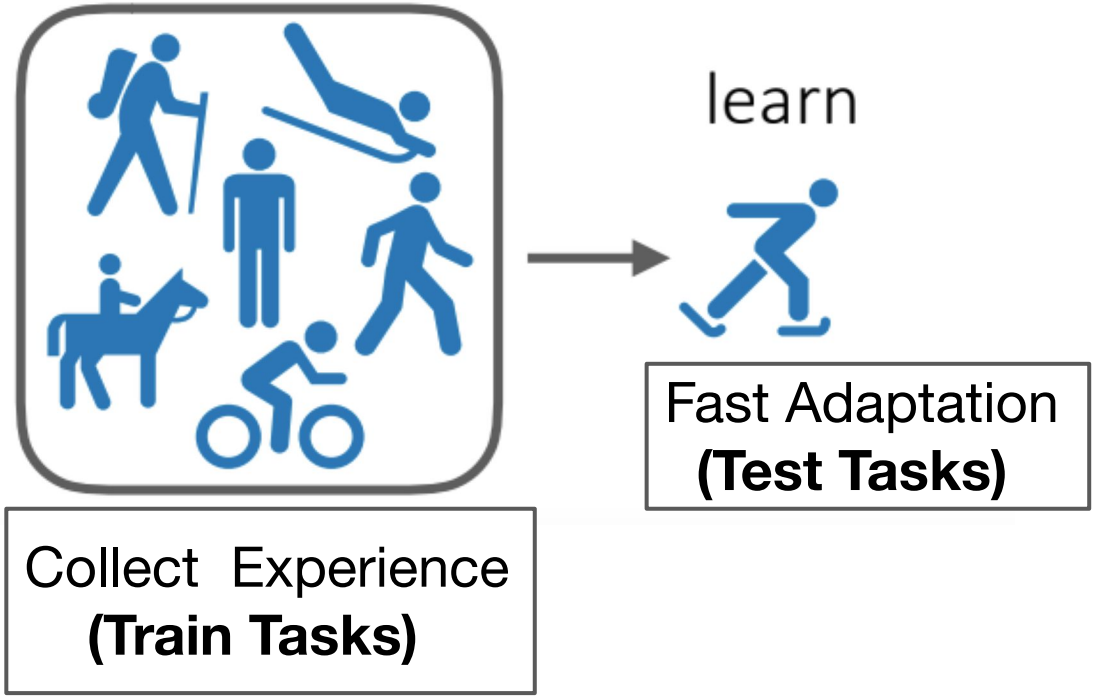






1 million timesteps

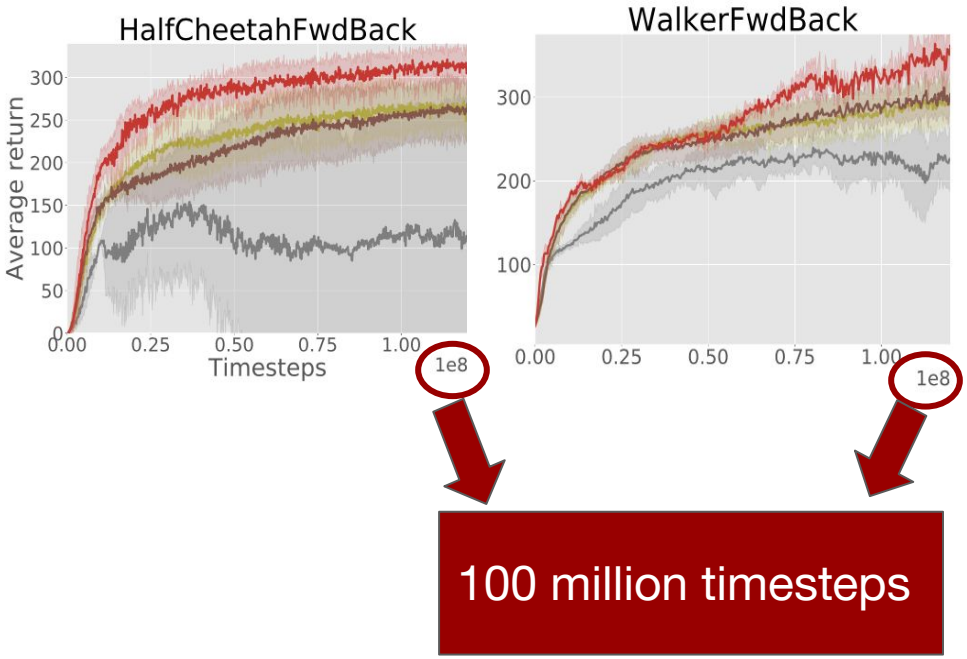
Meta-Learning



Challenges of Meta-training

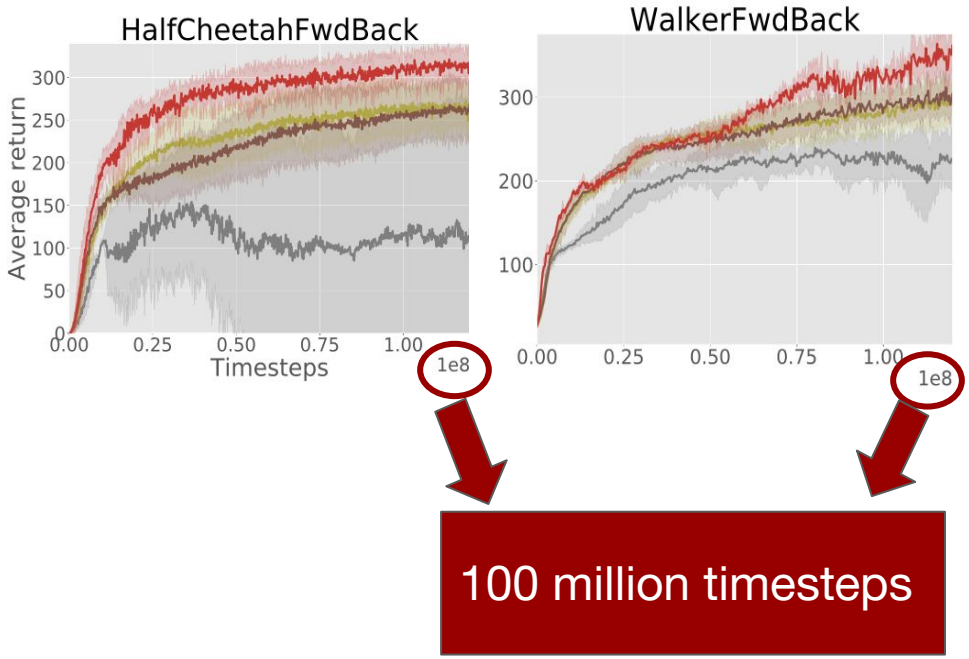
Challenges of Meta-training

High Sample Complexity



Challenges of Meta-training

High Sample Complexity



Harder Tasks (involving exploration / vision)

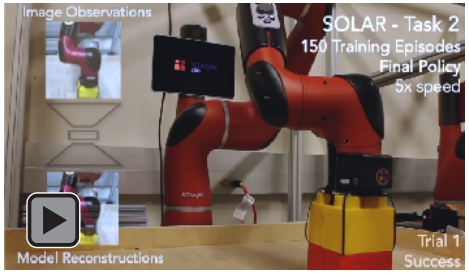


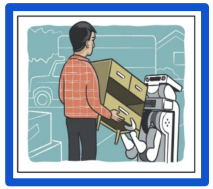
Image Observations

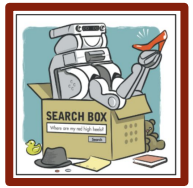
Sparse Reward

Guided Meta-Policy Search

Need a policy that can ***quickly*** adapt to solve any task from the distribution of training tasks

Train Set Tasks



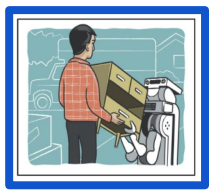


...

Guided Meta-Policy Search

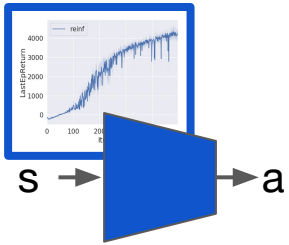
Need a policy that can ***quickly*** adapt to solve any task from the distribution of training tasks

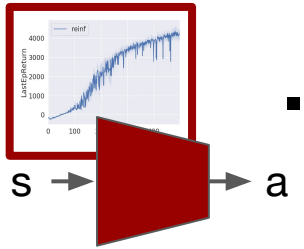
Train Set Tasks



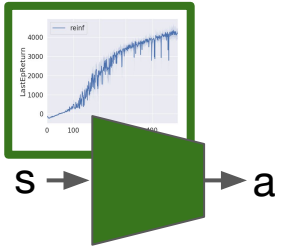
...

Learn **Local** Policies





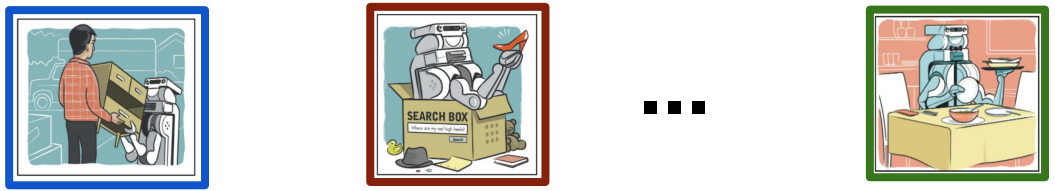
...



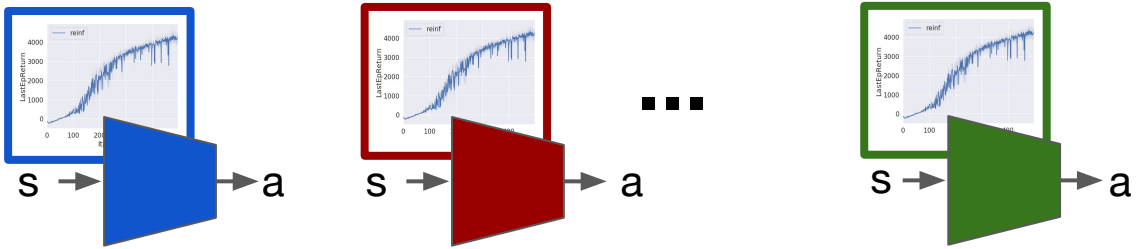
Guided Meta-Policy Search

Need a policy that can ***quickly*** adapt to solve any task from the distribution of training tasks

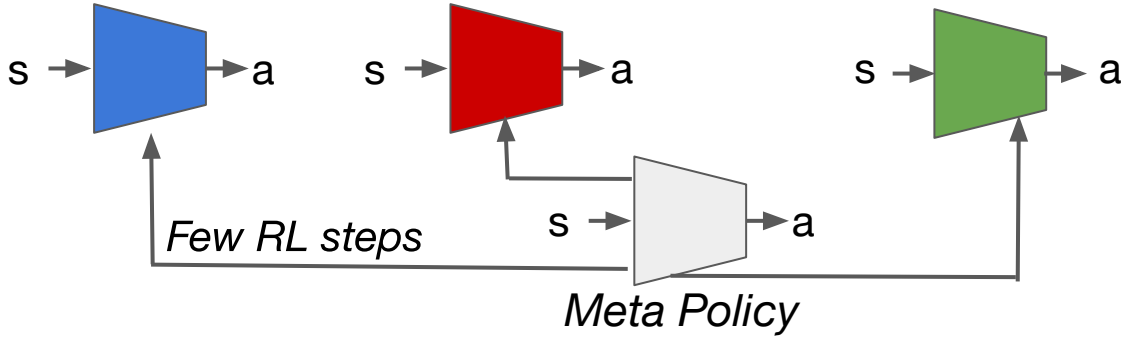
Train Set Tasks



Learn **Local** Policies



Updated Policies



Guided Meta-Policy Search

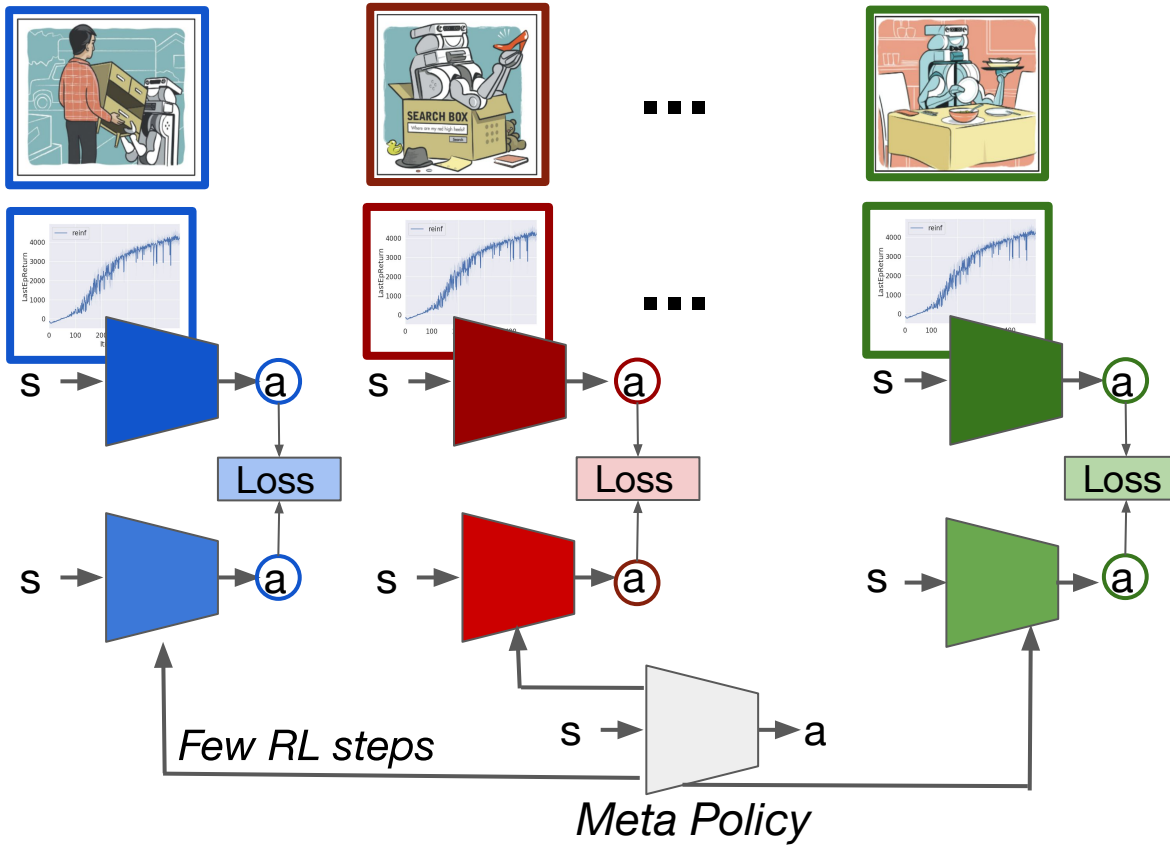
Need a policy that can ***quickly*** adapt to solve any task from the distribution of training tasks

Train Set Tasks

Learn **Local** Policies

Supervised Meta-RL Objective

Updated Policies



Guided Meta-Policy Search

Need a policy that can ***quickly*** adapt to solve any task from the distribution of training tasks

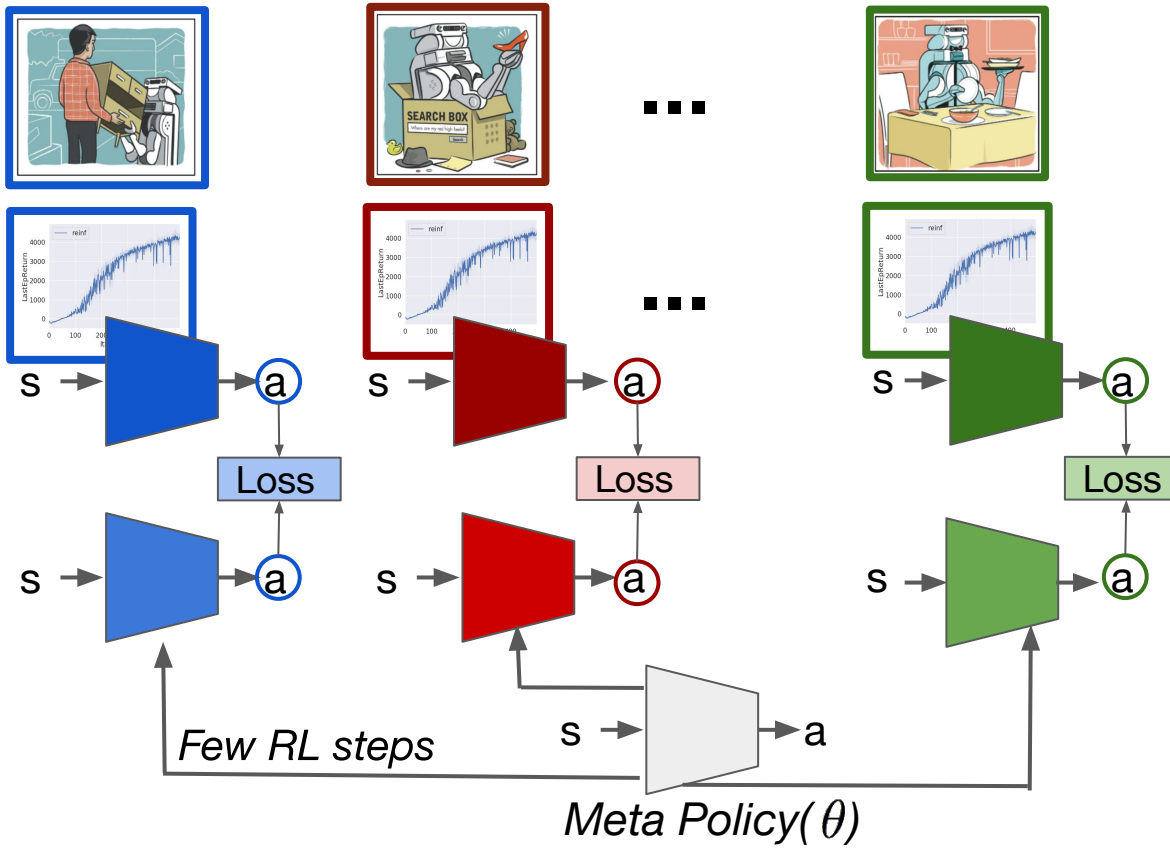
Train Set Tasks

Learn **Local** Policies

Supervised Meta-RL Objective

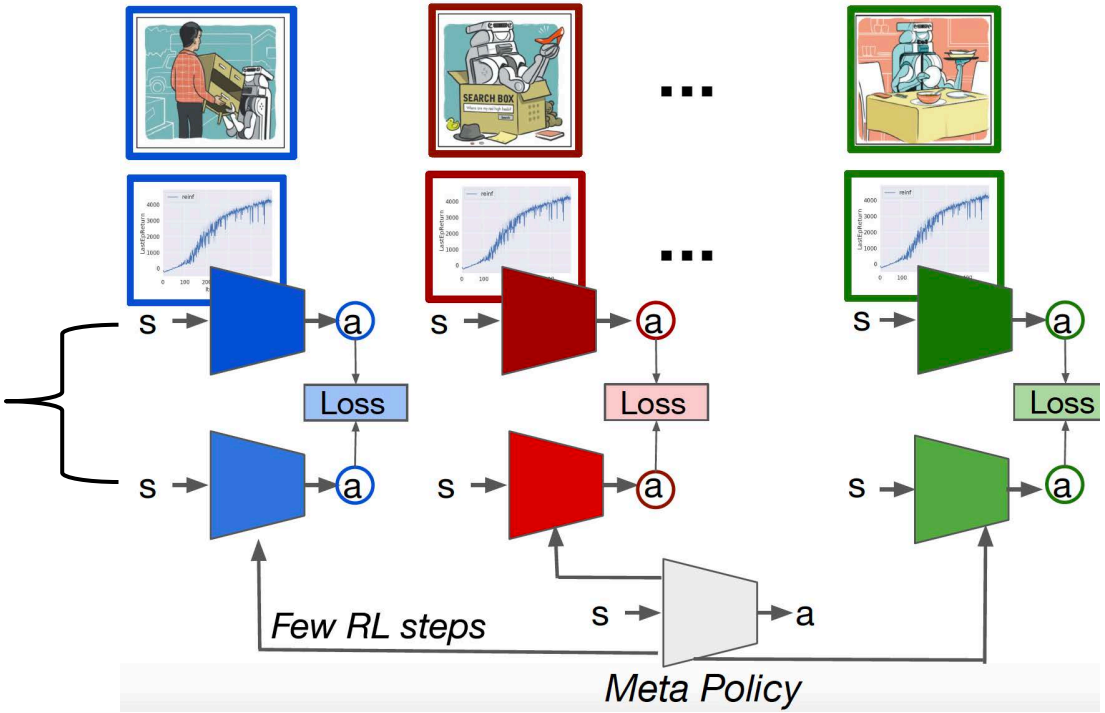
$$\max_{\theta} \sum_{\mathcal{T}_i} \sum_{(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{D}_{\mathcal{T}_i}} \log \pi_{\phi}(\mathbf{a}_t \mid \mathbf{s}_t)$$

Updated Policies (ϕ)



Guided Meta-Policy Search

Supervised Meta-RL Objective



- Off-policy learning
- Easier to optimize

Experimental Setup

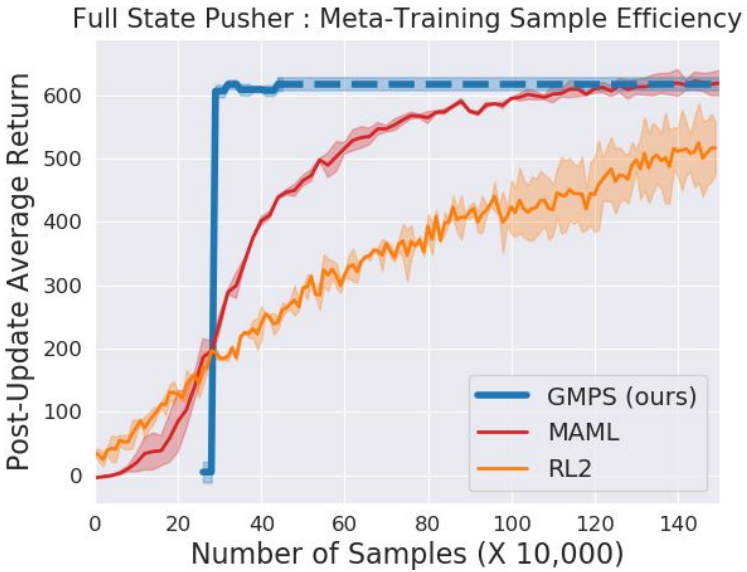
Sawyer: Pushing

Sawyer: Door Opening

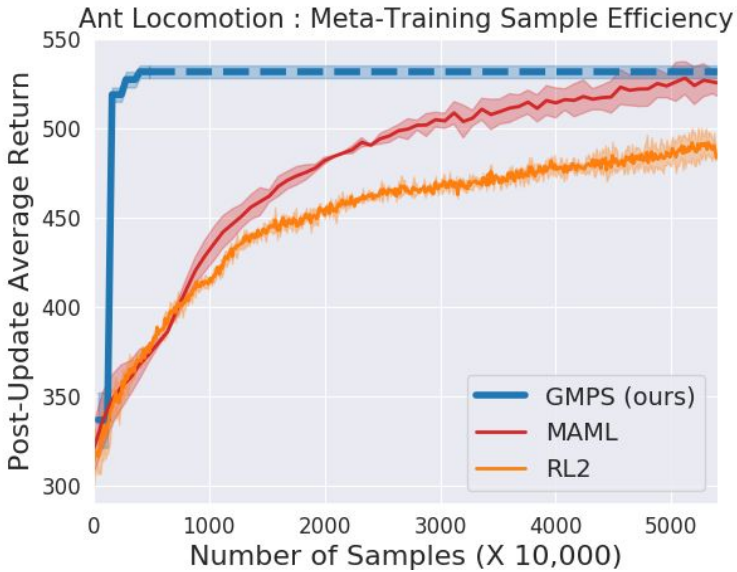
Legged Locomotion

Comparison to on-policy meta-RL methods (Sample Efficiency)

Sawyer Pushing



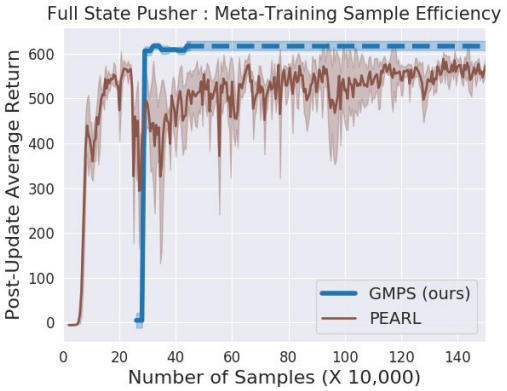
Ant Locomotion



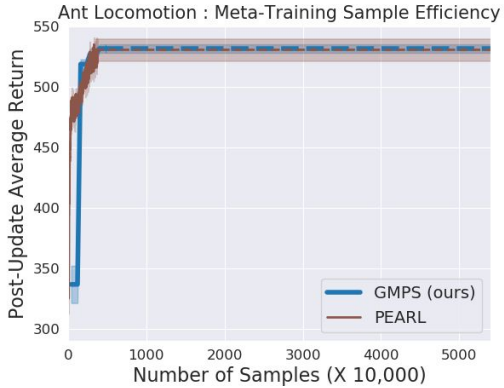
Comparison to off-policy meta-RL methods

Sample Efficiency

Sawyer Pushing



Legged Locomotion

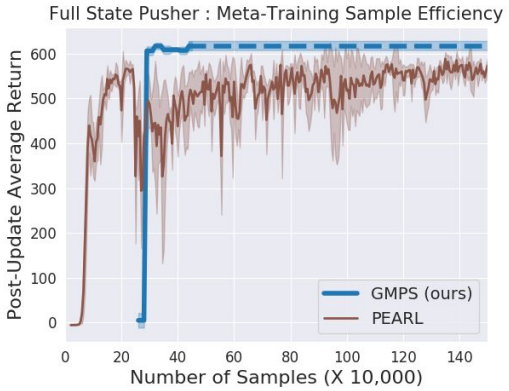


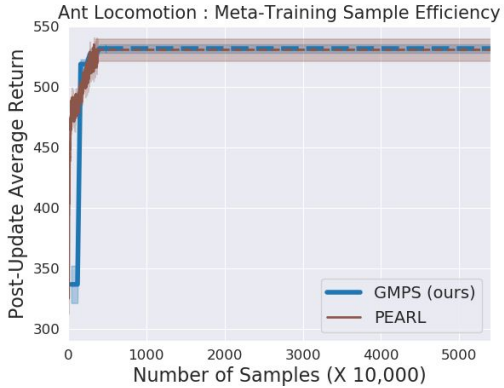
Comparison to off-policy meta-RL methods

Sawyer Pushing

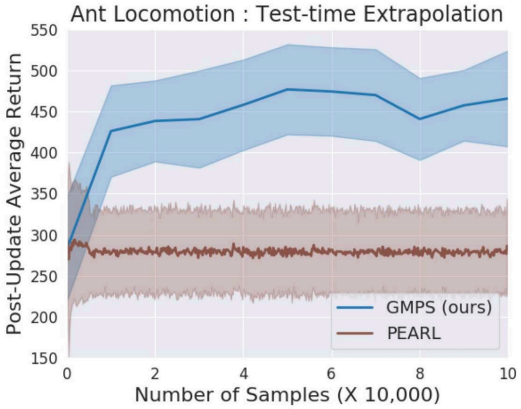
Legged Locomotion

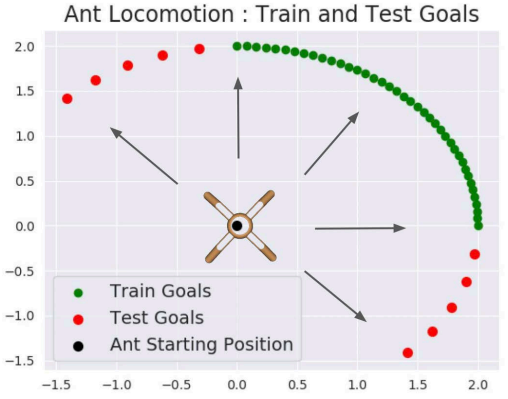
Sample Efficiency



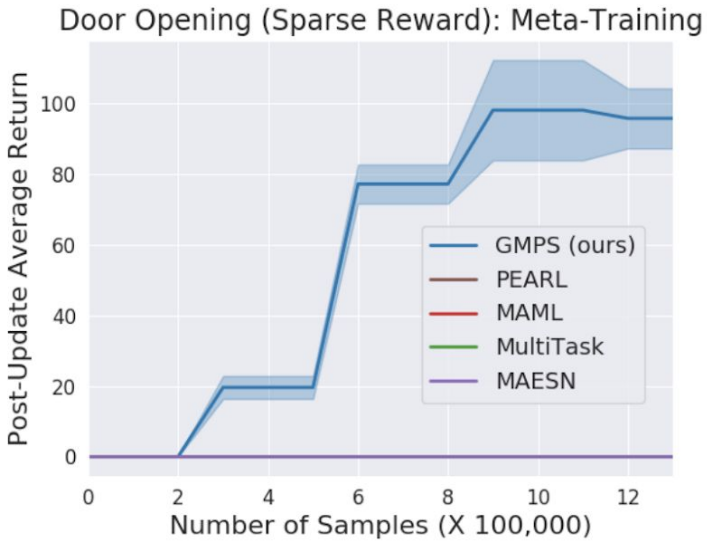


Extrapolation (Ant Locomotion)

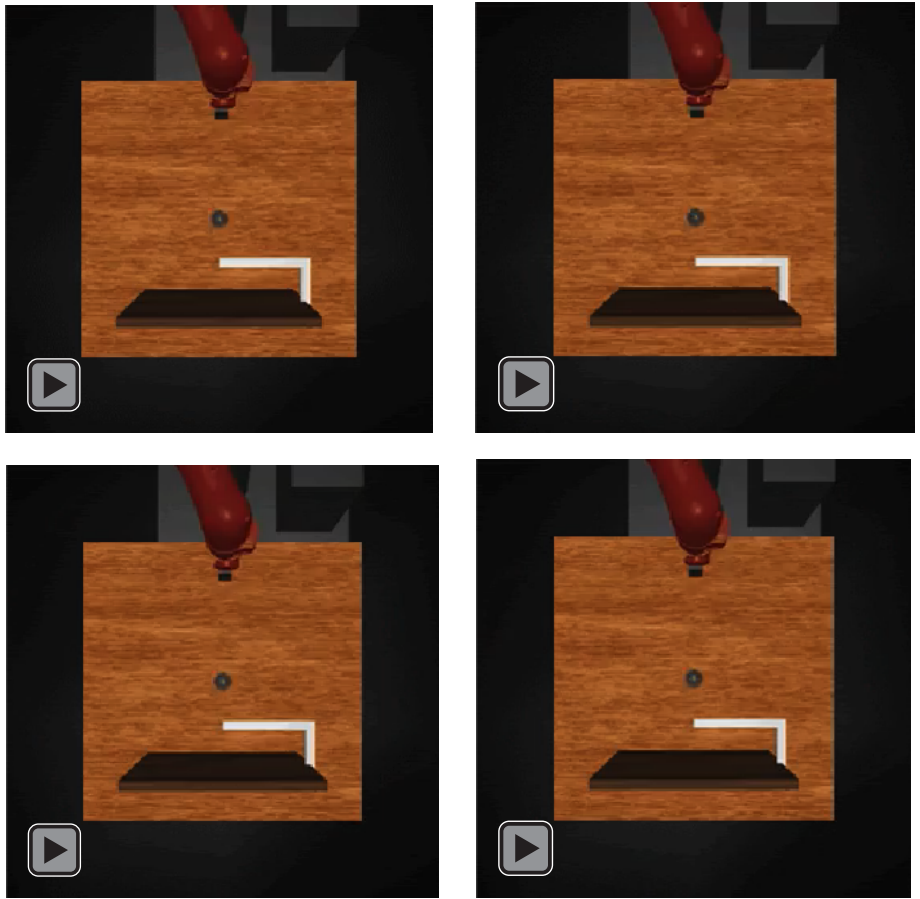




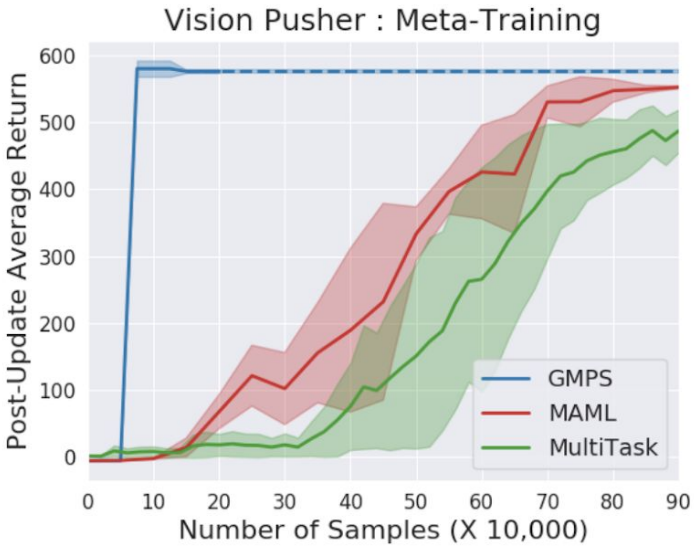
Meta-Learning from Demos : Sparse Reward



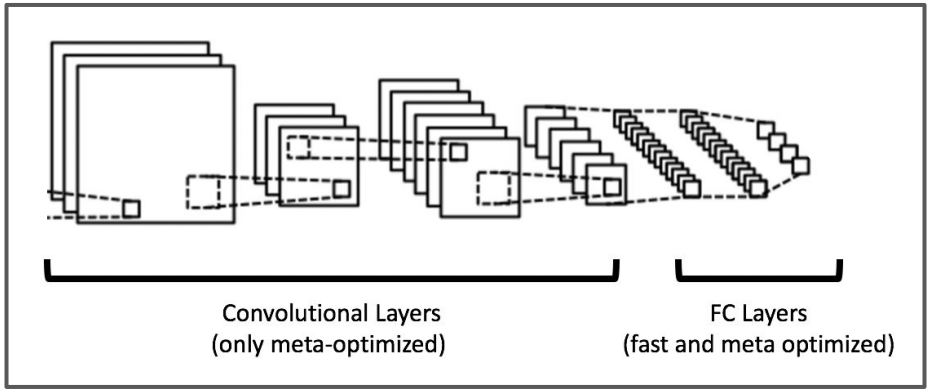
Door Opening

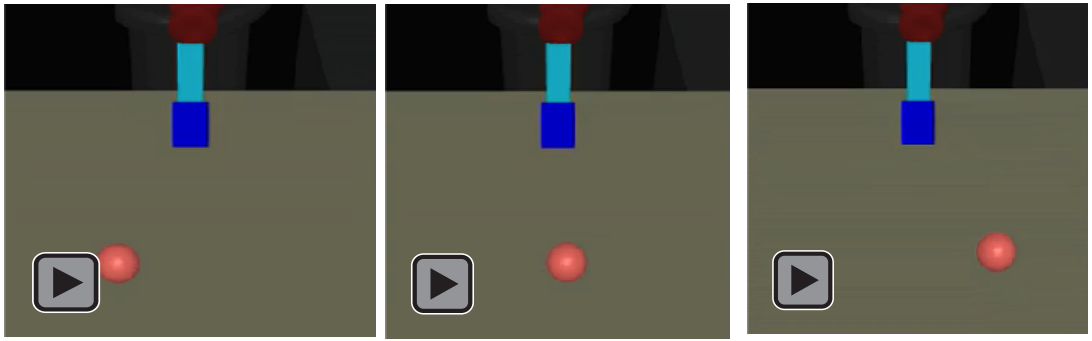


Meta-Learning from Demos : Visual Observations



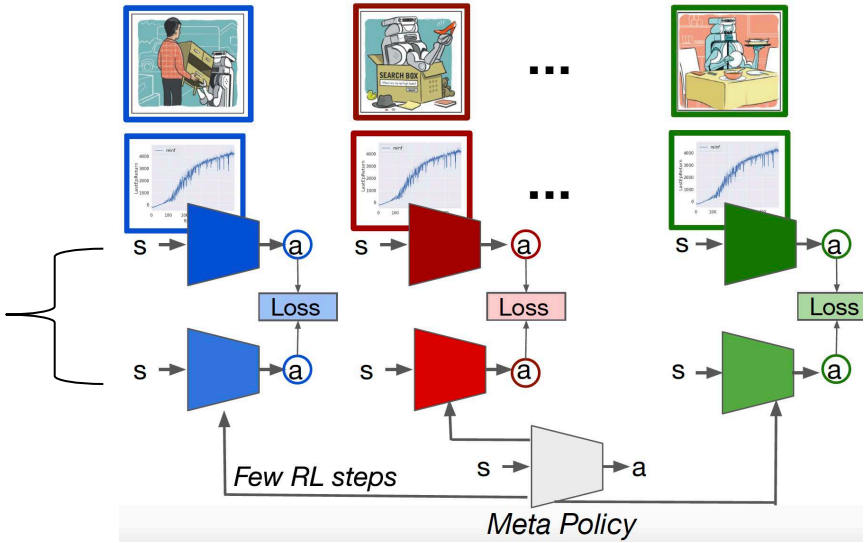
Pushing





GMPS **decouples** meta-optimization

Supervised
Meta-RL Objective



Please come visit our poster at **East Exhibition Hall B+C #42, 5:30 - 7:30 pm**

Github : russellmendonca/GMPS
Website : sites.google.com/berkeley.edu/guided-metapolicy-search
Contact : russellm@berkeley.edu

Thank You !

Abhishek Gupta

Rosen Kralev

Pieter Abbeel

Sergey Levine

Chelsea Finn

Please come visit our poster at **East Exhibition Hall B+C #42**

Github : [russellmendonca/GMPS]

Website : [sites.google.com/berkeley.edu/guided-metapolicy-search]

Contact : russellm@berkeley.edu