Learning a Prior over Intent via Meta-Inverse Reinforcement Learning

Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, Chelsea Finn
Where does the reward come from?

**Computer Games**

reward

Mnih et al. ‘15

**Real World Scenarios**

robotics
dialog
autonomous driving

what is the reward?
often use a proxy

*frequently easier to provide expert data*

**Inverse RL:** infer reward function from roll-outs of expert policy
Can we infer a reward from one or a few demonstrations?

Robots need **prior knowledge** & **context**.

How can robots **leverage prior experience** for **representing goals**?
Key intuition:
Learn a prior over human intent & then use learned prior to infer reward function in new scenario from a few demonstrations.

Navigation Problem:
- set of navigation tasks
- grass vs. dirt traversal preference
- landmark-directed navigation

Learn prior across tasks through meta-inverse reinforcement learning.
Meta-Inverse Reinforcement Learning

Meta-training time

Learn a prior over intent through meta-learning over meta-training tasks: $T_{\text{train}}$

Evaluation time

New task $T$

Rapid adaptation

$\theta' = \theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, T)$

Adapted reward $T\theta'$
Background: Model-Agnostic Meta-Learning

Key idea: Train over many tasks, to learn parameter vector $\theta$ that transfers

Fine-tuning

\[ \theta \leftarrow \theta - \alpha \nabla_\theta L_{\text{train}}(\theta) \]

Our method

\[ \min_\theta \sum_{\text{task } i} L^i_{\text{test}}(\theta - \alpha \nabla_\theta L^i_{\text{train}}(\theta)) \]

Intuition: Learning a prior over tasks, and at test time, inferring parameters under prior

(Grant et al. ICLR ’18)

Finn, Abbeel, Levine ICML ’17
Our approach: embed deep MaxEnt IRL [1,2] into meta-learning

\[ \min_{\theta} \sum_{\text{task } i} \mathcal{L}^i_{\text{test}}(\theta - \alpha \nabla_{\theta} \mathcal{L}^i_{\text{train}}(\theta)) \]


MandRIL
Meta Reward and Intention Learning
Experiments

At meta-test time:
- Provide a few demos

training environment  test environment (landmarks shuffled)

Comparisons:

MandRIL (ours) • Evaluate learned reward in original and new environment.

IRL from scratch • Compare value of optimal policy under true vs. learned reward

IRL from scratch • MaxEnt IRL only using demonstrations at meta-test time

Conditional Model • Condition reward model on visitation frequencies of demonstration

Recurrent Meta-Learner • Condition reward model on demonstration trajectories
Experiments

Meta-Test Training Performance

- MandRIL (ours)
- Conditional Model
- Recurrent Meta-Learner
- From Scratch

Meta-Test Testing Performance

- MandRIL (ours)
- Conditional Model
- Recurrent Meta-Learner
- From Scratch

value difference (lower is better)

number of demonstrations
Experiments

What about unseen landmarks?

Meta-Test Testing Performance

Unseen (Out of Domain) Objects

value difference (lower is better)

number of demonstrations
Future Directions

Do you need an entire demonstration to infer the goal?

Learn to **infer goals** from a few **positive examples**. (Xie, Singh, Levine, Finn ’18)

Explore less restricting IRL algorithms.

MaxEnt IRL applies to **tabular MDPs with known dynamics**. (so that it is easy to solve MDP in inner loop of IRL)
Reward learning is easier and more efficient with prior knowledge. Priors can be learned from data via meta-learning.
Reward learning is **easier** and **more efficient** with prior knowledge. Priors can be learned from data via **meta-learning**.

**Collaborators**

Kelvin Xu  
Ellis Ratner  
Anca Dragan  
Sergey Levine

**Questions?**

cbfinn@eecs.berkeley.edu