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

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: $\mathcal{T}_{\text{train}}$

Evaluation time

New task $\mathcal{T}$

Rapid adaptation

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

Adapted reward $\mathcal{T}\theta'$
**Key idea:** Train over many tasks, to learn parameter vector $\theta$ that transfers

**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

\[
\min_{\theta} \sum_{\text{task } i} A_i \log L_i^i \left( \mathcal{L}_{\text{train}}(\theta) \right)
\]

MandRIL
Meta Reward and Intention Learning

TODO: in future, make it clear that \( L \) is log probability. It's learning to do max likelihood.

At meta-test time:
- Provide a few demos

**Comparisons:**
- **MandRIL** (ours)
- **IRL from scratch**
- **Conditional Model**
- **Recurrent Meta-Learner**

**Experiments**
- Evaluate learned reward in original and new environment.
- Compare value of optimal policy under true vs. learned reward.
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?

![Graph showing meta-test testing performance and unseen (out of domain) objects.](image-url)
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