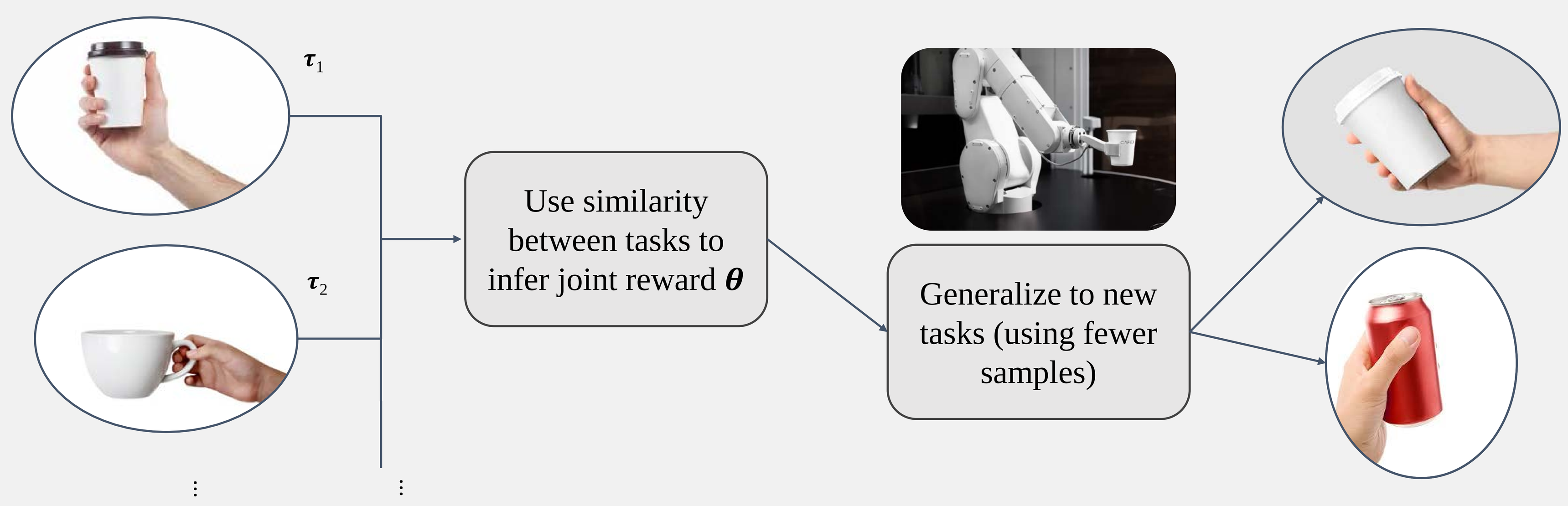
Multi-task Maximum Causal Entropy Inverse Reinforcement Learning

Adam Gleave & Oliver Habryka | {gleave,habryka}@berkeley.edu

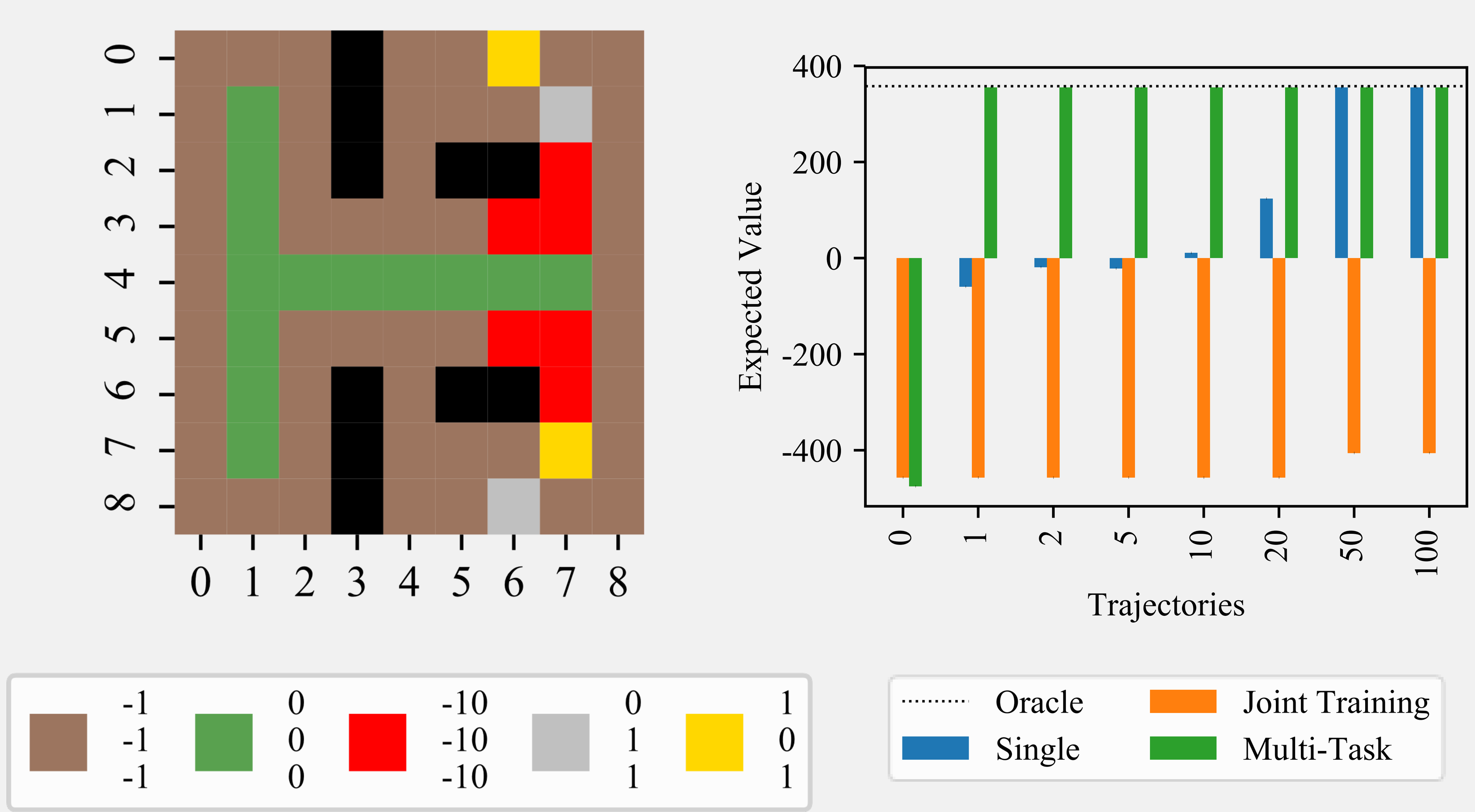
Introduction



Inverse reinforcement learning (IRL) is the task of determining the reward function that generated a set of trajectories: sequences of state-action pairs. Multi-task inverse IRL uses demonstrations of similar tasks, such as grasping different types of containers, to jointly infer the reward functions for each task. By exploiting the similarity between the reward functions, multi-task methods can achieve greater sample efficiency than conventional single-task IRL algorithms.

Previous work on the multi-task IRL problem builds on Bayesian IRL. Unfortunately, no extant Bayesian IRL methods scale to complex environments with high-dimensional, continuous state spaces such as robotics. By contrast, approaches based on maximum causal entropy (MCE) show more promise. Although the original MCE IRL algorithm is limited to discrete state spaces, recent extensions such as guided cost learning and adversarial IRL scale to challenging continuous control environment.

Regularized MCE IRL



The single-task IRL problem is to recover a reward function R given demonstrations τ from an MDP $M_i = (S, A, T, \gamma, \mu, R)$ and access to the world model (S, A, T, γ, μ). Maximum causal entropy (MCE) IRL assumes the reward function is linear in features over state-action pairs:

$$R(s,a) = \theta^T F(s,a).$$

For convenience, we write $F(\tau)$ to mean the (discounted) sum of features over state-action pairs in the trajectory. Maximum causal entropy IRL is equivalent to maximum causal likelihood estimation of θ given τ, i.e. finding θ that maximize the log likelihood:

$$\mathfrak{L}(\theta;\tau) = \sum_{t=0}^{T} \log \mathbb{P}(a_t \mid s_{0:t}\ , a_{0:t-1}).$$

The multi-task IRL problem is to jointly infer reward weights θ_i given expert demonstrations τ_i from MDPs $M_i = (S, A, T, \gamma, \mu, R_i)$. A useful inductive bias is that the rewards are similar between tasks, i.e. $\lambda \left\|\theta_i - \bar{\theta}\right\|^2$ should be small, where $\lambda > 0$. Letting π_i denote the softmax policy for reward parameters θ_i, the regularised loss and gradient are:

$$\mathfrak{L}(\theta_i;\tau) = \sum_{t=0}^{T} \log \mathbb{P}(a_t \mid s_{0:t}\ , a_{0:t-1}) + \tfrac{1}{2}\lambda \left\|\theta_i - \bar{\theta}\right\|^2,$$
$$\nabla\mathfrak{L}(\theta_i;\tau) = F(\tau_i) - F(\pi_i) - \lambda(\theta_i - \bar{\theta}).$$

We evaluate this multi-task IRL algorithm in a few-shot reward learning problem on the above gridworld. Each cell is either a wall (black), or one of five objects types. We define three different reward functions in terms of these object types, as specified the rows in the legend.

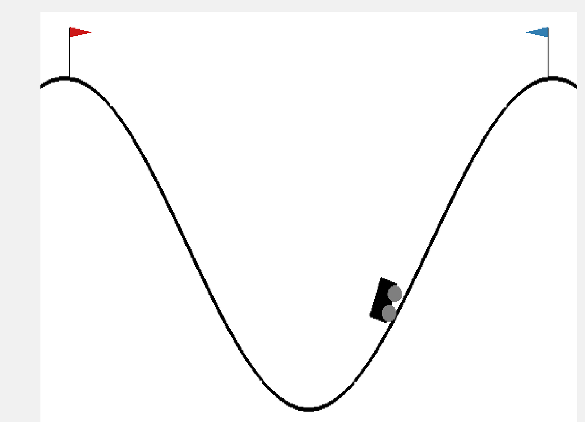
Our regularised MCE IRL algorithm is presented with a 1000 demonstrations from the middle and bottom reward functions, and between 1 and 100 trajectories from the top reward function. The value of the resulting policies (best of 5 seeds) is shown above, compared against two baselines and an optimal 'oracle' policy.

The regularised algorithm recovers a near-optimal policy after seeing only two trajectories, and for some seeds requires only a single trajectory. By contrast, the single-task baseline requires 50 or more trajectories to find a good policy, while the joint training baseline never succeeds.

Meta Adversarial IRL

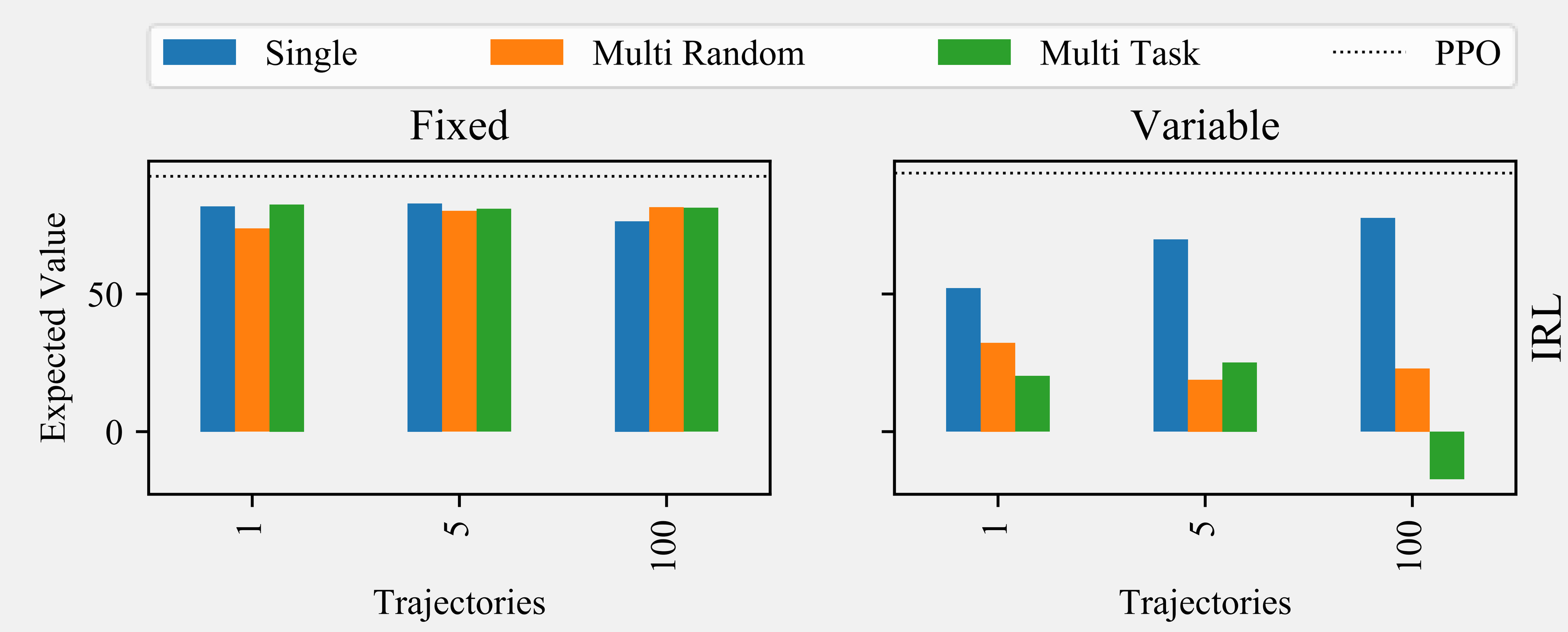
The MCE IRL algorithm has two major limitations: it assumes known dynamics T, and a linear reward function R. These shortcomings have been addressed by recent work such as adversarial IRL, a sample-based algorithm using a neural network to represent the reward R by parameters θ.

Meta-AIRL: Reptile and adversarial IRL (AIRL)
Randomly initialize network with parameters ϕ_0
for $t = 1 \dots T$:
  Sample task i with demonstrations τ_i
  Set $\theta_0 \leftarrow \phi_{t-1}$
  for $n = 1 \dots N$:
    $\theta_n \leftarrow \mathrm{AIRL}(\theta_{n-1}, \tau_i)$, one step of AIRL
  Update $\phi_t \leftarrow \phi_{t-1} + \alpha(\theta_N - \phi_{t-1})$



We used Reptile, a computationally efficient meta-learning algorithm, to find an initialisation ϕ of the reward network that can be quickly finetuned for a new task. Our algorithm, above, repeatedly samples a task and then runs N steps of adversarial IRL starting from the current initialisation ϕ. The initialisation is then updated along the line to the final iterate of adversarial IRL.

We evaluate on a multi-task variant of the mountain car continuous control problem, illustrated above. The episode ends as soon as the car touches either flag. One flag is the *goal* and gives 100 reward, the other a *decoy* with a -100 penalty. We create two test cases, called *fixed and variable*, each consisting of two environments. In the fixed test case, the side of the goal flag is static. In the variable test case, the colour of goal flag is static, but the side varies between episodes.



In the fixed test case (left), both the single-task baseline and our meta-AIRL algorithm produce near-optimal solutions. We conjecture this is because the optimal policy is unimodal, making it simple to extrapolate from a single trajectory. In the variable test case (right), single-task AIRL fails to find a good solution even after observing 100 trajectories. Reptile can only learn a good initialisation in the outer loop when progress is made in the AIRL inner loop, so unsurprisingly our meta-AIRL algorithm also fails. Note the variable test case has a bimodal expert policy.

Our findings suggest that adversarial IRL succeeds only in environments with a unimodal optimal policy. In such environments, a handful of trajectories is sufficient to learn the reward, leaving little room for improvement from using meta-learning. However, many practical tasks (such as multi-step assembly) have multimodal expert policies, making this a pressing area for further research.

Conclusion & Further Work

Sample efficient solutions to the multi-task IRL problem are critical for enabling real-world applications, where collecting human demonstrations is expensive and slow. The multi-task IRL problem has previously been studied exclusively from a Bayesian IRL perspective. In this paper we took the alternative approach of formulating the multi-task problem inside the maximum causal entropy IRL framework.

Our first contribution uses the original MCE IRL algorithm, by adding a regularisation term to the loss. Experiments find our regularized MCE IRL algorithm can perform one-shot imitation learning in an environment that otherwise requires hundreds of demonstrations to learn.

In preliminary work, we combined the Reptile meta-learner with adversarial IRL, a sample-based MCE IRL algorithm. Testing revealed that adversarial IRL can only learn from unimodal expert policies, seriously limiting the applicability of meta-AIRL. We conjecture this limitation in adversarial IRL is related to the mode collapse in generative adversarial networks (GAN). A fruitful research direction might be to apply recent innovations in GAN training, such as unrolling the optimisation of the discriminator or variational learning, to stabilise adversarial IRL training.

Further Information

Full paper: bit.ly/MultiTaskIRL
Source code: bit.ly/MultiTaskIRLCode