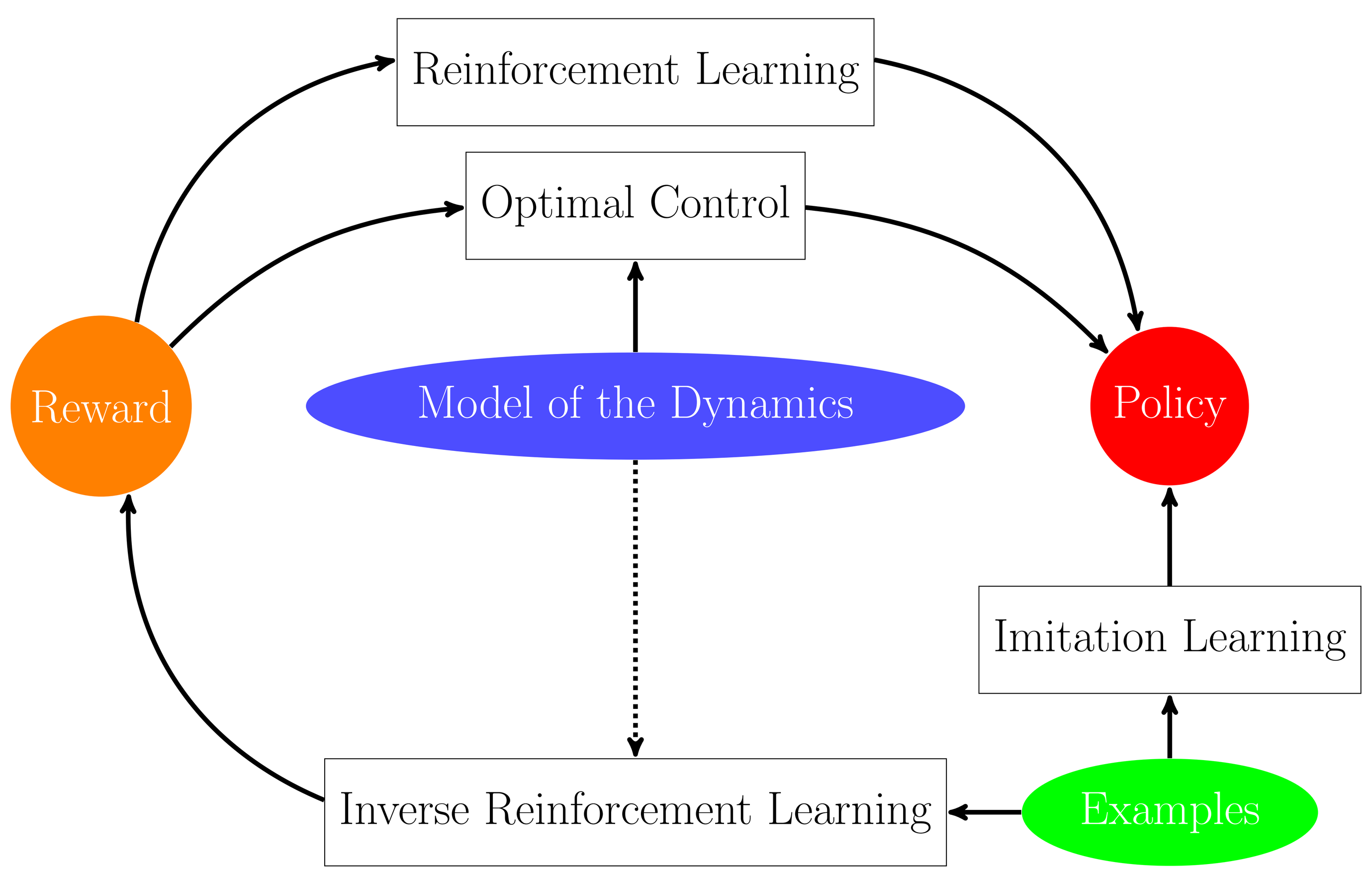
Relative Entropy Inverse Reinforcement Learning

Abdeslam Boularias Jens Kober Jan Peters

Max Planck Institute for Intelligent Systems, Tübingen, Germany

1 Overview



We consider the problem of inverse reinforcement learning when a model of the dynamics is unavailable.

2 Background

2.1 Markov Decision Process (MDP)

A Markov Decision Process is a tuple $(\mathcal{S}, \mathcal{A}, T, R)$, where

- $\mathcal{S}$ is a set of states,
- $\mathcal{A}$ is a set of actions,
- T are transition probabilities with $T(s,a,s') = P(s'|s,a)$ for $s, s' \in \mathcal{S}, a \in A$,
- R is a reward function where $R(s,a)$ is the reward given for action a in state s.

2.2 Policies

A policy π is a function that maps every state into an action or a distribution on the actions. The expected average reward received by following a policy π is given by

$$J(\pi) = \frac{1}{h}\mathrm{E}_{s_t,a_t}\Big[\sum_{t=0}^{h-1} R(s_t,a_t)\Big|d_0,\pi,T\Big],$$

where d_0 is the initial state distribution and h is the horizon.

2.3 Inverse Reinforcement Learning (IRL)

- Designing a reward function for matching a complex behavior can be a challenging problem. It is often easier to provide examples of the desired behavior [1].
- IRL consists in learning a reward function that explains an observed behavior.
- The reward is assumed to be a linear function of state-action features f_i,

$$R(s,a) = \sum_{i=0}^{k-1} \theta_i f_i(s,a).$$

- The learned reward function, parameterized by θ, is used to generalize the observed behavior.

3 Relative Entropy Inverse Reinforcement Learning

- A trajectory of states and actions $s_1a_1,\dots s_ha_h$ is denoted by τ.
- The average value of feature f_i along a trajectory τ is denoted by $f_i(\tau)$.
- The empirical average of feature f_i in the observed trajectories is denoted by $\hat{f}_i$
- Let Q be a basline distribution on the trajectories. Q can be uniform (Maximum Entropy [2]), or an initial approximation of the observed behavior.
- Find a distribution P that is as close as possible to Q, while each feature has an average under P that is close to its average in the observed trajectories.

3.1 Problem statement

Relative Entropy IRL is formulated as the problem of minimizing the relative entropy between P and Q,

$$\min_P \sum_{\tau\in\mathcal{T}} P(\tau)\ln\frac{P(\tau)}{Q(\tau)},$$

subject to the following constraints

$$\forall i \in \{1,\dots,k\} : \Big|\sum_{\tau\in\mathcal{T}} P(\tau)f_i(\tau) - \hat{f}_i\Big| \le \epsilon_i, \quad (1)$$

$$\sum_{\tau\in\mathcal{T}} P(\tau) = 1, \quad (2)$$

$$\forall \tau \in \mathcal{T} : P(\tau) \ge 0. \quad (3)$$

3.2 Solution

The subgradient of the dual function g is given by

$$\frac{\partial}{\partial\theta_i} g(\theta) = \hat{f}_i - \sum_{\tau\in\mathcal{T}} P(\tau|\theta) f_i(\tau) - \alpha_i\epsilon_i,$$

where $\alpha_i = 1$ if $\theta_i \ge 0$ and $\alpha_i = -1$ otherwise.

The parameterized trajectory distribution P is given by

$$P(\tau|\theta) = \frac{1}{Z(\theta)} Q(\tau)\exp\Big(\sum_{i=1}^{k}\theta_i f_i(\tau)\Big).$$

The probability $Q(\tau)$ is given by $d(\tau)u(\tau)$, where $d(\tau)$ is the joint probability of the state transitions in τ, and $u(\tau)$ is the joint probability of the actions conditioned on the states in τ.

✗ The subgradient of the dual function cannot be calculated if $d(\tau)$ is unknown.

3.3 Stochastic subgradient with Importance Sampling

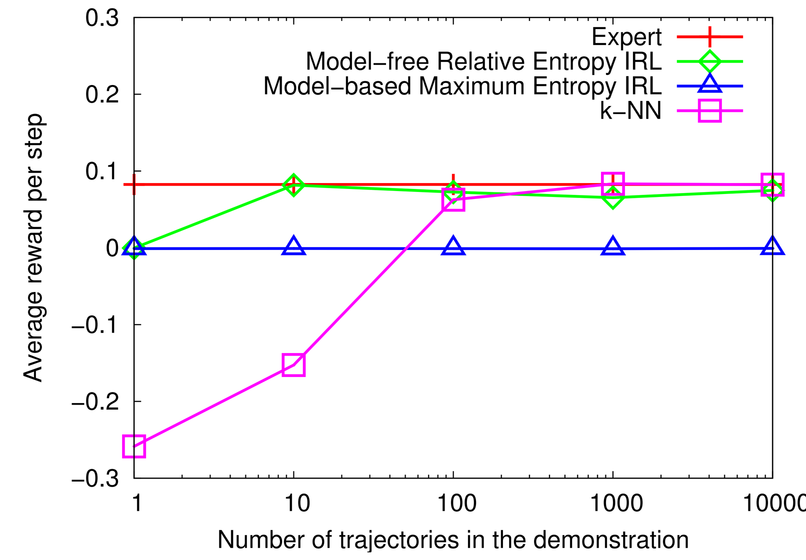
Let $\pi(\tau)$ be the joint probability of the actions in τ under a sampling policy.

$$\frac{\partial \hat{g}}{\partial\theta_i}(\theta) = \hat{f}_i - \frac{\sum_\tau \frac{u(\tau)}{\pi(\tau)}\exp\big(\sum_i\theta_i f_i(\tau)\big) f_i(\tau)}{\sum_\tau \frac{u(\tau)}{\pi(\tau)}\exp\big(\sum_i\theta_i f_i(\tau)\big)} - \alpha_i\epsilon_i.$$

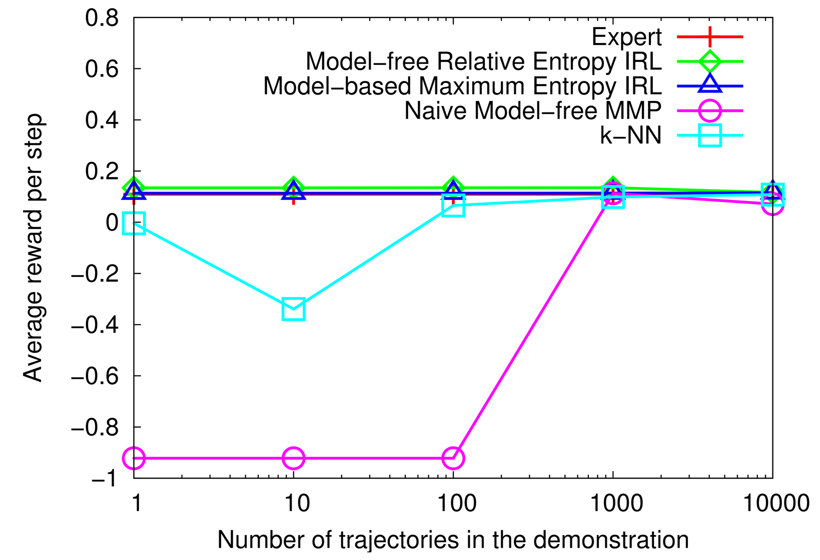
4 Experiments

4.1 Gridworld and Racetrack

We compare the Relative Entropy IRL to: (1) the model-based Maximum Entropy IRL [2] where a model is estimated from the demonstrated trajectories, (2) a naive model-free variant of Maximum Margin Planning (MMP [3]) where the forward problem is solved by reinforcement learning, (3) a supervised learning approach (k-NN).

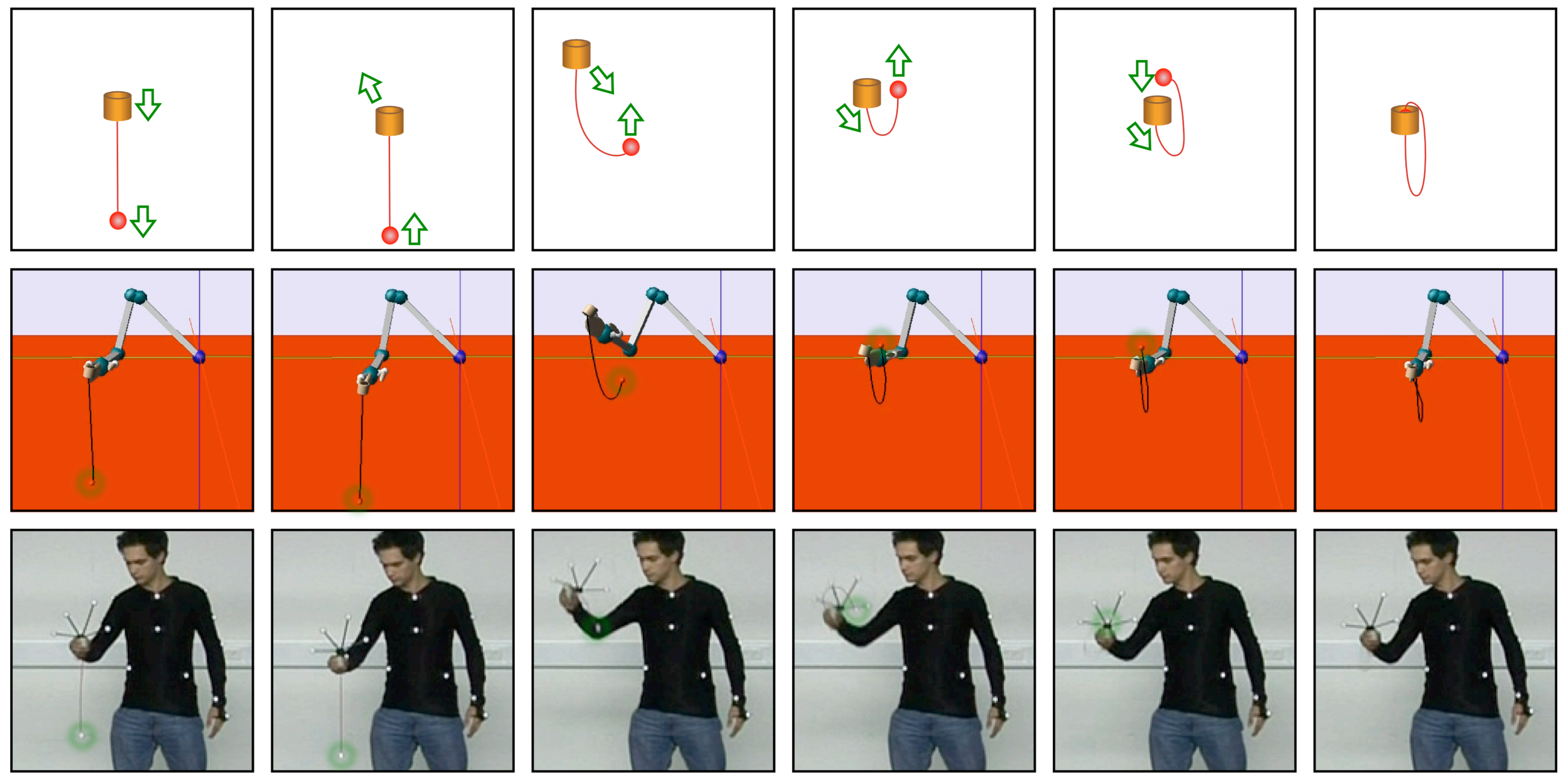


Gridworld

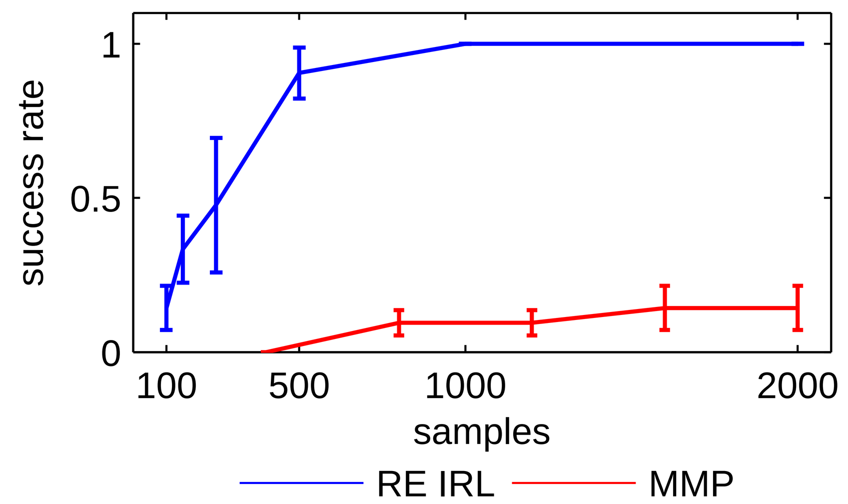


Racetrack

4.2 Ball-in-a-cup



The figure above shows schematic drawings of the Ball-in-a-Cup motion, the final learned robot motion as well as a motion-captured human motion. We used 17 trajectories provided by a human expert. The figure on the right shows the average success rate as a function of the number of sampled trajectories that were used for learning the reward function.



References

[1] Pieter Abbeel and Andrew Ng. Apprenticeship Learning via Inverse Reinforcement Learning. ICML 2004.

[2] Ziebart, B., Maas, A., Bagnell, A., and Dey, A. Maximum Entropy Inverse Reinforcement Learning. AAAI 2008.

[3] Nathan Ratliff, J. Andrew Bagnell and Martin Zinkevich. Maximum Margin Planning. ICML 2006.

AISTATS 2011, Ft. Lauderdale, April 2011