Modelling Human Decision-making based on Aggregate Observation Data

Antti Kangasrääsiö
Samuel Kaski
first.last@aalto.fi

Problem Setting

We wish to infer the task, preferences or limitations of users when they are performing complex decision tasks

We use a Reinforcement Learning agent to model the behavior of the user (ie. we define a parametric environment and a task)

Given observations of the user's behavior, we wish to infer the parameters of the task and the environment
→ Inverse Reinforcement Learning

New: We assume that the granularity of the observations is large
→ only aggregate observations of user behavior
→ very common problem setting in practice, but no methods exist

Our Contributions

We propose an extension of the IRL problem, called Inverse Reinforcement Learning from Summary Data (IRL-SD)

We derive a Bayesian likelihood for this problem, but demonstrate that it may be very expensive to evaluate

We propose an approximate ABC-likelihood that is faster to evaluate

We propose a BO method for performing inference

IRL-SD Problem

Let M be a MDP (S, A, T, R, γ) with parameters θ. Let the true parameters be $\theta^* \in \Theta$ and assume agent behaving according to an optimal policy for M_{θ^*}. Assume the agent has taken paths $(\xi_1, \ldots, \xi_N)$ and we observe *summaries* $\Xi_\sigma = (\xi_{1\sigma}, \ldots, \xi_{N\sigma})$, where $\xi_{i\sigma} \sim \sigma(\xi_i)$ and σ is a known summary function. The *inverse reinforcement learning problem from summary data (IRL-SD)* is then:

Given (1) set of summaries Ξ_σ of an agent demonstrating optimal behavior; (2) summary function σ; (3) MDP M with θ unknown; (4) bounded space Θ; and optionally (5) prior $P(\theta)$.
Estimate $\hat{\theta} \in \Theta$ such that simulated behavior from $M_{\hat{\theta}}$ agrees with Ξ_σ, or the posterior $P(\theta|\Xi_\sigma)$.

Exact Likelihood

Assume both $|S|$ and $|A|$ are finite and that the maximum number of actions that can be performed within an observed episode is T_{max}. Denote the finite set of all plausible trajectories by $\Xi_{ap} \subseteq S^{T_{max}+1} \times A^{T_{max}}$.

The likelihood for θ given $\Xi_\sigma = (\xi_{1\sigma}, \ldots, \xi_{N\sigma})$ is now

$$L(\theta|\Xi_\sigma) = \prod_{i=1}^{N} \left[P(\xi_{i\sigma}|\theta)\right] = \prod_{i=1}^{N} \Big[\sum_{\xi_i \in \Xi_{ap}} \left[P(\xi_{i\sigma}|\xi_i) P(\xi_i|\theta)\right] \Big],$$

where

$$P(\xi_{i\sigma}|\xi_i) = P\big(\sigma(\xi_i) = \xi_{i\sigma}\big),$$

and

$$P(\xi_i|\theta) = P(s_0^i) \prod_{t=0}^{T_i - 1} \left[\pi_\theta^*(s_t^i, a_t^i) P(s_{t+1}^i | s_t^i, a_t^i)\right].$$

Approximate Likelihood

Assume a function for generating summary datasets Ξ_σ^{sim} given MDP M, parameters θ, number of episodes N, and summary function σ: $\mathtt{RLSUM}(M_\theta, N, \sigma)$. Also assume a *discrepancy function* δ,

$$\delta(\Xi_\sigma^A, \Xi_\sigma^B) \to [0, \infty),$$

which quantifies the dissimilarity between two observation datasets.

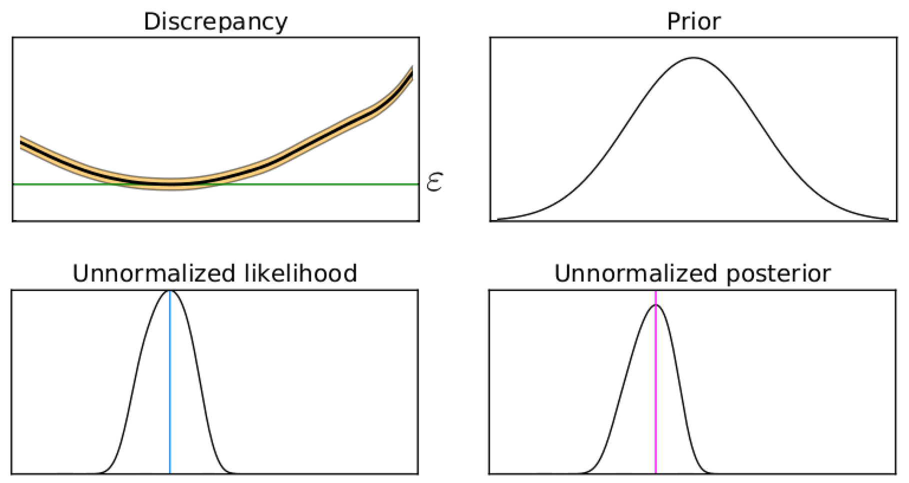
By combining $\mathtt{RLSUM}(M_\theta, |\Xi_\sigma|, \sigma)$ with δ, we define

$$d_\theta \sim \delta(\mathtt{RLSUM}(M_\theta, |\Xi_\sigma|, \sigma), \Xi_\sigma).$$

The distribution of d_θ corresponds with the ability of θ to satisfy our requirements for solving the IRL-SD problem. Finally we define an approximate likelihood function,

$$\tilde{L}_\varepsilon(\theta|\Xi_\sigma) = P(d_\theta \le \varepsilon | \theta),$$

where the approximation threshold $\varepsilon \in [0, \infty)$.



Inference Algorithms

Algorithm 1 Exact Maximum Likelihood Inference Algorithm for IRL-SD

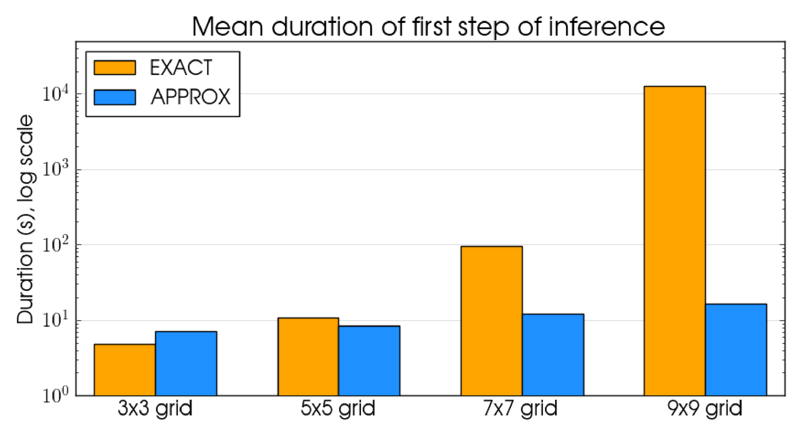
Input: $M, \Xi_\sigma, \Theta, H, N_{opt}$
Output: $\hat{\theta}_{ML}$
$D \leftarrow \varnothing$
for $i = 1$ **to** N_{opt} **do**
  $\theta_i \leftarrow \arg\max_\theta Acq(\theta|D, H)$
  $\pi_{\theta_i}^* \leftarrow \mathtt{RL}(M_{\theta_i})$
  $l_\theta \leftarrow -\log L(\theta_i|\Xi_\sigma)$
  $D \leftarrow \{D, (\theta_i, l_\theta)\}$
end for
$\hat{\theta}_{ML} \leftarrow \arg\min_\theta G_\mu(\theta|D, H)$

Algorithm 2 Approximate Maximum Likelihood Inference Algorithm for IRL-SD

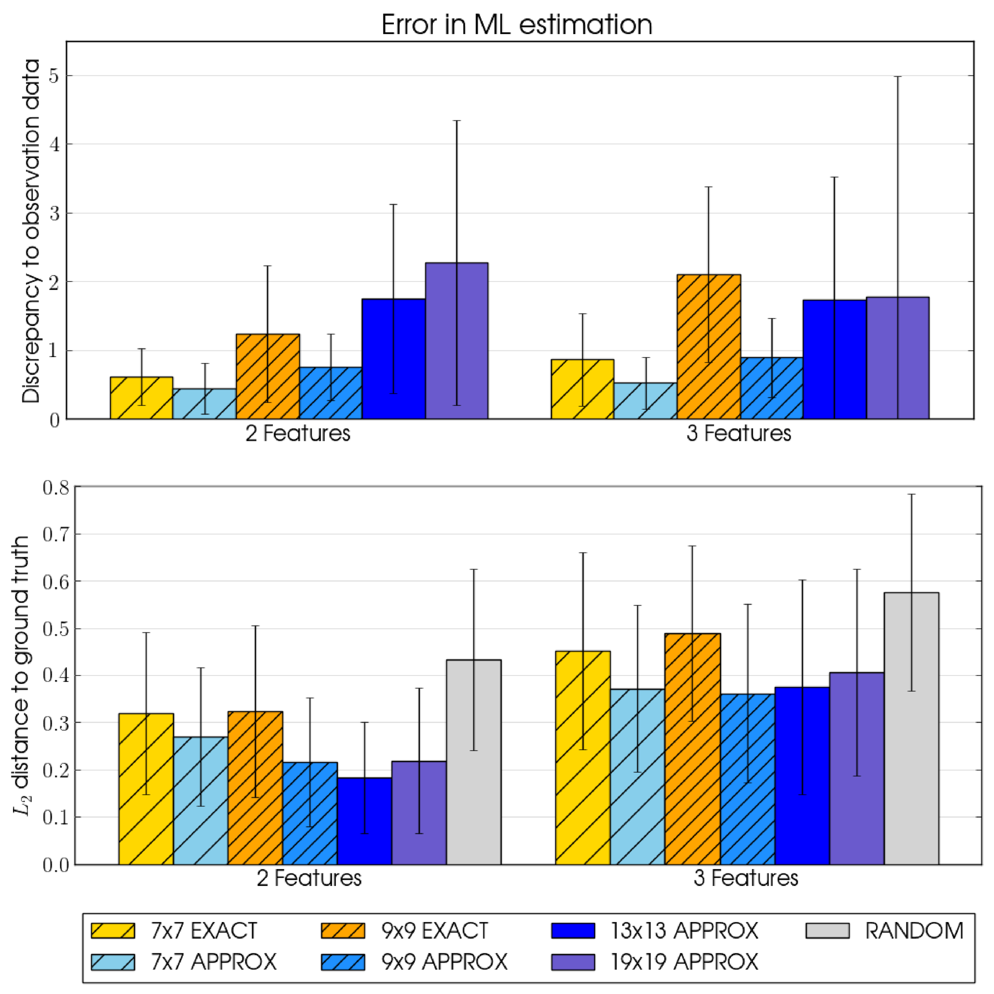
Input: $M, \Xi_\sigma, \Theta, H, N_{opt}$
Output: $\hat{\theta}_{ML}$
$D \leftarrow \varnothing$
for $i = 1$ **to** N_{opt} **do**
  $\theta_i \leftarrow \arg\max_\theta Acq(\theta|D, H)$
  $\Xi_\sigma^{sim} \leftarrow \mathtt{RLSUM}(M_{\theta_i})$
  $d_\theta \leftarrow \delta(\Xi_\sigma^{sim}, \Xi_\sigma)$
  $D \leftarrow \{D, (\theta_i, d_\theta)\}$
end for
$\hat{\theta}_{ML} \leftarrow \arg\min_\theta G_\mu(\theta|D, H)$

Experiments

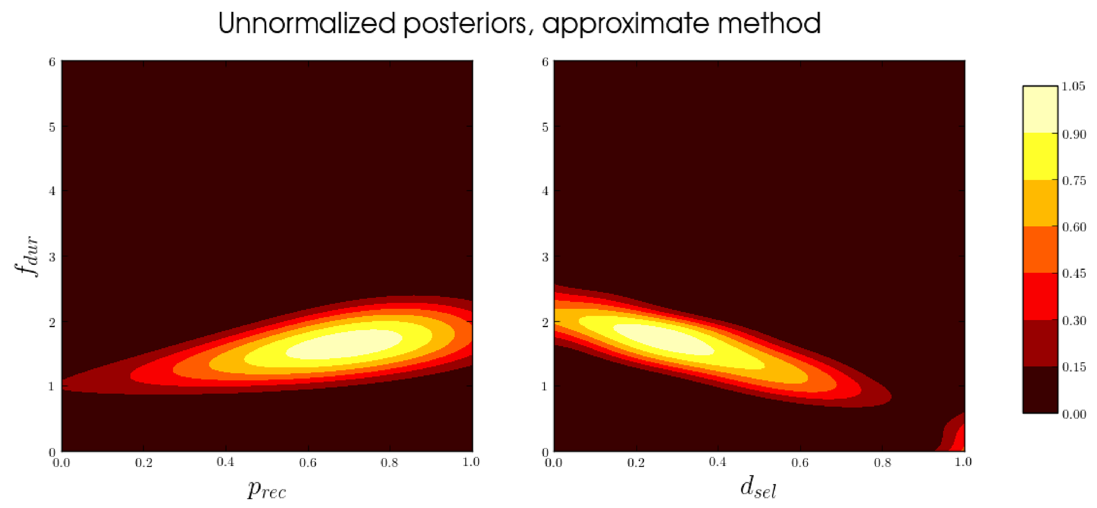
Grid World
Algorithm runtime (one step)



Grid World
Inference quality



Visual search in drop-down menus
Full posterior estimation



Conclusion
Regarding partial observability in IRL, there now exists formulations for three different situations:
(1) Agent has partial observability of the environment state → POMDP model
(2) External observer has partial observability on *state* level → traditional IRL methods can be extended
New: (3) External observer has partial observability on *path* level → presented methods for IRL-SD can be used