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 (i.e. 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 ($S, A, T, R, \gamma$) with parameters $\theta$. Let the true parameters be $\theta^* \in \Theta$ and assume agent behaviour according to an optimal policy for $M_{\theta^*}$. Assume the agent has taken paths $\{z_1, \ldots, z_N\}$ and we observe summaries $Z = \{z_1, \ldots, z_N\}$, where $z_i = \sigma(z_i)$ and $\sigma$ is a known summary function. The inverse reinforcement learning problem from summary data (IRL-SD) is then:

Given (1) set of summaries $Z$ of an agent demonstrating optimal behavior; (2) summary function $\sigma$; (3) MDP $M$ with $\theta$ unknown; (4) bounded space $\Theta$; and optionally (5) prior $P(\theta)$

Estimate $\theta \in \Theta$ such that simulated behavior from $M_{\theta}$ agrees with $Z$, or the posterior $P(\theta|Z)$

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 $Z \subseteq \mathbb{S}^{T_{max}} \times A^{T_{max}}$

The likelihood for $\theta$ given $Z = \{z_1, \ldots, z_N\}$ is now

$L(\theta|Z) = \prod_{i=1}^N P(z_i|\theta) = \prod_{i=1}^N \left[ \sum_{z_i \in Z} P(z_i|\theta) P(\theta|z_i) \right]$

where $P(z_i|\theta) = P(\sigma(z_i) = z_i)$,
and $P(\theta|z_i) = P(\theta|z_i) \prod_{t=0}^{T_{max}} [\pi_t^{\theta}(s,t) \pi_t^{\theta}(s',t)]$

Approximate Likelihood
Assume a function for generating summary datasets $Z^{sum}$ given MDP $M$, parameters $\theta$, number of episodes $N$, and summary function $\sigma$: $\text{RLSTM}(N, M_{\theta})$. Also assume a discrepancy function $\delta$, $\delta(z^{sum}, z) \rightarrow [0, \infty)$

which quantifies the dissimilarity between two observation datasets.

By combining $\text{RLSTM}(N, M_{\theta})$ with $\delta$, we define $d_{\theta} = \delta(\text{RLSTM}(N, M_{\theta}), Z)$

The distribution of $d_{\theta}$ corresponds with the ability of $\theta$ to satisfy our requirements for solving the IRL-SD problem. Finally, we define an approximate likelihood function,

$L_{\delta}(\theta|Z) = P(d_{\theta} \leq \epsilon|\theta)$

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

Inference Algorithms

Algorithm 1: Exact Maximum Likelihood Inference Algorithm for IRL-SD
Input: $M, Z, \Theta, H, N_{opt}$
Output: $d_{opt}$

for $i = 1$ to $N_{opt}$ do

$\theta_i \sim \text{arg max}_{\theta} L(\theta|Z)$

end for

Algorithm 2: Approximate Maximum Likelihood Inference Algorithm for IRL-SD
Input: $M, Z, \Theta, H, N_{opt}$
Output: $d_{opt}$

for $i = 1$ to $N_{opt}$ do

$\theta_i \sim \text{arg max}_{\theta} L_{\delta}(\theta|Z)$

end for

Experiments

Grid World
Algorithm runtime (one step)

Mean duration of training (s)

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
(3) External observer has partial observability on path level → presented methods for IRL-SD can be used