We wish to infer the parameters of the task, preferences or limitations of users when they are performing tasks involving strategic behavior in inverse reinforcement learning (IRL) context. A limitation with existing methods for IRL is that they make very specific assumptions about the type of observation data: trajectories denoted as \( \xi = (s_0, a_1, s_1, \ldots, a_T, s_T) \). We extend this setting to arbitrary noise models \( \sigma(\xi) \).

**Background:** In IRL, a RL model is used to explain the strategic behavior of a user in a situation similar to a Markov decision process (S, A, T, R, \( \gamma \)). The strategic behavior is assumed to follow an optimal policy for the MDP. Given observations of the user’s behavior, we wish to infer the parameters of the MDP, such that the optimal policy matches the observed behavior.

**Problem Setting**

We wish to infer the parameters of the task, preferences or limitations of users when they are performing tasks involving strategic behavior in inverse reinforcement learning (IRL) context. A limitation with existing methods for IRL is that they make very specific assumptions about the type of observation data: trajectories denoted as \( \xi = (s_0, a_1, s_1, \ldots, a_T, s_T) \). We extend this setting to arbitrary noise models \( \sigma(\xi) \).

**Background:** In IRL, a RL model is used to explain the strategic behavior of a user in a situation similar to a Markov decision process (S, A, T, R, \( \gamma \)). The strategic behavior is assumed to follow an optimal policy for the MDP. Given observations of the user’s behavior, we wish to infer the parameters of the MDP, such that the optimal policy matches the observed behavior.

**Our Contributions**

We demonstrate that IRL inference is possible even when the observation noise model is an arbitrary function \( \sigma(\xi) \). This extends the state of the art which is only able to deal with few special types of observation noise (missing / probabilistic state observations).

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

We propose two approximations: a Monte-Carlo estimate and an ABC estimate, which are faster to evaluate.

**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 behaving according to an optimal policy for \( M^\theta \). Assume the agent has taken paths \( (\xi_1, \ldots, \xi_n) \) and we observe summaries \( Z = (Z_1, \ldots, Z_n) \), where \( Z_i = \sigma(\xi_i) \) and \( \sigma \) is a known summary function. The inverse reinforcement learning problem from summary data (IRL-SD) is then:

**Given** (1) a 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**

\[
L(\theta|Z) = \prod_{i=1}^{N} P(\xi_i|\theta) = \prod_{i=1}^{N} \sum_{\xi_i} P(\xi_i|\theta) P(\xi_i),
\]

\[
P(\xi_i|\theta) = P(\xi_0) \prod_{t=1}^{T-1} \sigma_\theta(a_t|s_t) P(s_{t+1}|s_t, a_t).
\]

**Monte-Carlo Likelihood**

\[
\hat{L}(\theta|Z) = \frac{1}{N_{MC}} \sum_{i=1}^{N} P(\xi_i|\theta) P(\xi_i|\theta)
\]

\[
P(\xi_i|\theta) = P(\xi_0) \prod_{t=1}^{T-1} \sigma_\theta(a_t|s_t) P(s_{t+1}|s_t, a_t).
\]

Expensive to evaluate

Applicable when we know \( \sigma \) as a distribution

**ABC Likelihood**

Assume a function for generating summary datasets \( Z_i \) given MDP \( M \), parameters \( \theta \), number of episodes \( N \), and summary function \( \sigma \): \( RLSUM(M, N, \sigma) \). Also assume a discrepancy function \( d \), which quantifies the dissimilarity between two observation datasets.

By combining \( RLSUM(M, N, \sigma) \) with \( \theta \), we define

\[ d_{\theta} \sim d(RLSUM(M, N, \xi), \xi). \]

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

\[
\hat{L}(\theta|Z) = P(\theta|\xi_0, \xi_1, \ldots, \xi_T) = \prod_{i=1}^{N} \sum_{d_i} P(d_{\theta} \leq d_i|\xi_i),
\]

where the approximation threshold is \( d \in [0, \infty) \).

**Inference**

GP surrogate fit using Bayesian optimization

**Experiments (Grid World)**

Algorithm runtime (one step) with different grid sizes

**Inference quality (error to ground truth, prediction error) with different grid sizes and dimensionality of reward function

**Full Posterior Inference**

Model: Visual search in drop-down menus

**Summary**

Regarding partial observability in IRL, there now exists formulations for three different situations:

1. Agent has partial observability of the environment state \( \rightarrow \) POMDP model
2. External observer has partial observability on state level \( \rightarrow \) traditional IRL methods can be extended
3. External observer has partial observability on path level \( \rightarrow \) presented methods for IRL-SD can be used