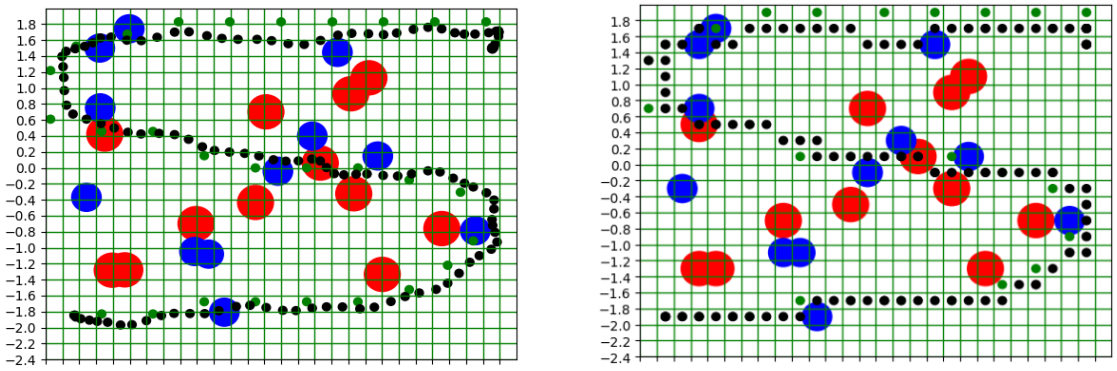
Human Reward Learning with BIRL

October 27, 2017

1 Domain Representations



(a) Original Data (b) Discretized Data

Figure 1: Example of data discretization

In order to make the problem tractable by BIRL, the test area is discretized into a 2D gridworld of size 28×22 with 0.2×0.2 m^2 cells. Each cell is a state in the MDP. There are 8 directions/actions an agent can take at any state to move to an adjacent state. The (center) location of targets, obstacles, pathpoints and the subject's waypoints are approximated with their closest discrete state as shown in figure 1. The problem is formulated as weight-learning for the three different features: targets, obstacles and pathpoints.

The three features are represented using three different continuous values at each state. More specifically, the nine cells around targets and obstacles all have a feature value of 1.0 for the corresponding feature, while the the centers of pathpoints take an increasing value as their order increases. The reward at any given state is computed as the linear combination of these features using their corresponding weights. The observations are a set of state-action pairs extracted from the human's trajectory.

2 Experimental Design Choices

Since the discretization as well as test-time error can introduce noise into the data, only a (randomly sampled) subset of available human state-action pairs is used per itration of BIRL. The model with lowest training error across 50 iterations is used as its final prediction.

The parameters for BIRL are set empirically. The confidence factor α is set at 100 and the chain length is set to be 3000 (since there are only three values, i.e. feature weights to be tweaked, which is relatively small). 0.85 is used as the discount factor for MDPs.

3 Results