Multi-Robot Inverse Reinforcement Learning Under Occlusion with State Transition Estimation

(Extended Abstract)

Kenneth Bogert
THINC Lab, University of Georgia
Athens, GA 30602
kbogert@uga.edu

Prashant Doshi
THINC Lab, University of Georgia
Athens, GA 30602
pdosh@cs.uga.edu

ABSTRACT

Multi-robot inverse reinforcement learning (mIRL) is broadly useful for learning, from passive observations, the behaviors of multiple robots executing fixed trajectories and interacting with each other. In this paper, we relax a crucial assumption in IRL to make it better suited for wider robotic applications: we allow the transition functions of other robots to be stochastic and do not assume that the transition *error* probabilities are known to the learner. Challenged by occlusion where large portions of others' state spaces are fully hidden, we present a new approach that maps stochastic transitions to distributions over features.

Categories and Subject Descriptors

I.2.9 [**Robotics**]: Workcell organization and planning

General Terms

Algorithms; Performance

Keywords

inverse reinforcement; machine learning; multi-robot systems

1. INTRODUCTION

We study an application setting involving two mobile robots independently executing simple cyclic trajectories for perimeter patrolling. Both robots' patrolling motions are disturbed when they approach each other in narrow corridors leading to an interaction. A subject robot observes them from a *hidden* vantage point that affords partial observability of their trajectories only. It's task is to penetrate the patrols and reach a goal location without being spotted. Thus, its eventual actions do not impact the other robots.

Inverse reinforcement learning (IRL) [3, 5] is well suited as a starting point here because the task is to learn the preferences of passively-observed experts from their state-action trajectories. Previously, Bogert and Doshi [2] models each observed robot in the setting as guided by a policy from a Markov decision process (MDP) and utilizes IRL generalized for occlusion. However, the interactions between the patrollers must be modeled as well. As these are sparse and scattered, the robots are modeled as playing a game at each point of interaction. Consequently, this method labeled mIRL*+Int generalizes IRL – so far limited to single-expert contexts – to multiple experts exhibiting sparse interactions and whose trajectories are partially occluded from the learner.

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

Key model assumptions of popular IRL methods are that the expert's stochastic transition function is completely known to the learner as in IRL for apprenticeship learning [1] and in Bayesian IRL [4]. Alternately, the transition function is effectively deterministic and thus is easily approximated from the observed trajectories [5] with the assumption that transition randomness has a limited effect on the learner's final behavior. The prior knowledge requirement is often difficult to satisfy in practice, for example, in scenarios that are not cooperative such as the patrolling application. Alternately, the supposed impotency of transition errors is a strong assumption in the context of robots.

We partially relax IRL's prior knowledge requirements and tread a middle path: we limit to those settings where a mobile robot's stochastic transition function may be viewed as composed of a deterministic core perturbed by transition error probabilities that make it stochastic. Given a state-action pair, the learner knows the intended next state of each expert. However, the transition error probabilities are unknown. Of course, the learner may learn the complete transition functions using supervised learning if it observes the experts fully and long enough. But, partial occlusion and a finite observation time motivate sophisticated methods.

Challenged by occlusion, we present mIRL$^*_{/T}$+Int a novel method based on the key insight that different transitions share underlying component features, and features associated with observed state-action pairs may transfer information to transitions in occluded portions. Subsequently, mIRL$^*_{/T}$+Int maps each state-action pair to a feature subset. Thus, probability of success of an action in a state resulting in the intended next state is the joint probability of success of all features involved in that action. Our task reduces to finding the probability of success of each feature from observations that do not inform each feature but instead pertain to *feature aggregates*.

2. LEARNING OTHERS' TRANSITIONS

Let ψ: $S \times A \rightarrow S$ map an observed robot's transition from state, s, given action a to a particular next state, s'. The function, ψ, gives the intended outcome of each action from each state. We may view this as a *deterministic* transition function. Of course, actions may not always generate their intended outcomes leading to small errors in the corresponding transitions. Furthermore, parts of the robot's trajectory may be occluded from the subject robot, and the robot may be guided by a policy. Both these factors make it unlikely that the learning robot will observe every action in every state enough times to reliably compute the full transition function. *Therefore, we focus on learning the probability of transitioning to the intended state given a state-action pair for an observed robot I,* $T_I(s, a, \psi(s, a))$. The remaining probability mass, $1 - T_I(s, a, \psi(s, a))$, could be distributed uniformly among the states that are the intended outcomes of other actions given the state, or

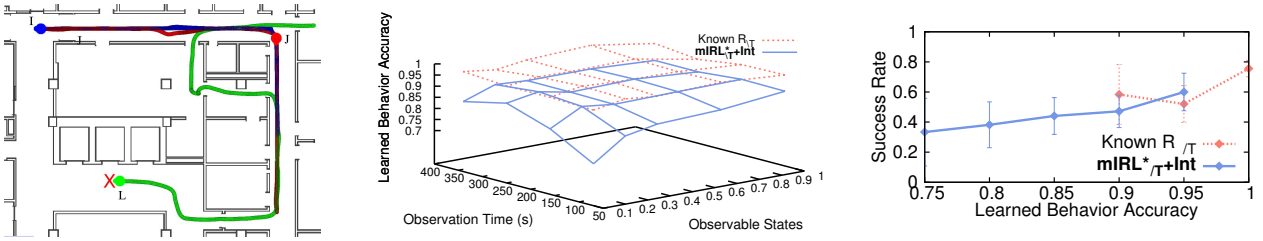

Figure 1: ***(left)*** **Hallways of a building patrolled by I (in blue) and J (in red) with the start location of L inside a room looking out of an open door. The goal location is marked with an 'X'.** ***(middle)*** **Learned behavior accuracy of mIRL$^*_{/T}$+Int and Known R$_{/T}$ for different occlusion rates and observing times.** ***(right)*** **Improving accuracy of learned behavior correlates almost linearly with success rate.**

wholly assigned to a default error state. This approach requires that ψ is available to the learner (but not the probability with which $\psi(s,a)$ results). In order to learn robustly under occlusion, our approach is based on the following key observation: If transition probabilities are a function of underlying component outcome probabilities, then the observed trajectory may inform associated component probabilities. Subsequently, if some of these components are shared with transitions in occluded states, then information is transferred that facilitates obtaining occluded transition probabilities.

We begin by mapping each state-action to a subset of lower-level transition features. Let $\xi_I^{s,a} = \{\tau_1, \ldots, \tau_k\}$ be the subset of independent features mapped to a state-action pair, $\langle s,a \rangle$, where each feature, $\tau \in \mathcal{T}_I$, is a binary random variable whose states are τ and $\bar{\tau}$. Subsequently, define for a transition, $\langle s,a,\psi(s,a) \rangle$,

$$T_I(s,a,\psi(s,a)) = Pr(\tau_1, \tau_2, \ldots, \tau_{|\xi_I^{s,a}|}) \approx \prod\nolimits_{\tau \in \xi_I^{s,a}} Pr(\tau)$$

The equation above casts the problem of inversely learning the transition function as the problem of learning the distributions of the state-action features. *However, the challenge is that we may not be able to pinpoint the performance of the various features in the observed trajectory; rather we obtain* **aggregated** *empirical distributions.* An observed trajectory of length T is a sequence of state-action pairs, $\{\langle s,a \rangle^0, \langle s,a \rangle^1, \ldots, \langle s,\phi \rangle^T\}$, where ϕ is the null action. From this, we obtain the probabilities of transitioning to the intended state given the previous state and action, denoted by $q_I^{\psi(s,a)}$, as simply the proportion of times the intended state is observed as the next state in the trajectory. Notice that the probability, $q_I^{\psi(s,a)}$, obtained from an observed trajectory is equivalent to $T_I(s,a,\psi(s,a))$. Consequently,

$$\prod\nolimits_{\tau \in \xi_I^{s,a}} Pr(\tau) = q_I^{\psi(s,a)} \quad (1)$$

While $\xi_I^{s,a}$ tells us which features are assigned to each state-action and Eq. 1 constrains the feature distributions, *we arrive at an ill-posed problem where there could be many feature distributions satisfying observed transition probabilities that serve as aggregates*.

One way to make progress in an underconstrained problem is to utilize the principle of maximum entropy optimization [5] because it makes the least assumptions beyond the problem formulation. In this context, mIRL$^*_{/T}$+Int maximizes the sum total entropy of all feature distributions. Constraints for this nonlinear optimization problem are given by Eq. 1 for each state-action pair present in the observed trajectory. Previously unseen actions could have been performed in the occluded portions of other robot's trajectory. Nevertheless, these actions map to feature variables in $\mathcal{T}_I$. *As some of the features in $\mathcal{T}_I$ are factors in observed actions, we may obtain (partially) informed transition distributions for the unseen actions as well under the maximum entropy principle.* mIRL$^*_{/T}$+Int solves the nonlinear optimization to obtain feature distributions.

3. PERFORMANCE EVALUATION

We evaluate mIRL$^*_{/T}$+Int in the domain introduced by Bogert and Doshi [2] and discussed in Fig. 1. L utilizes the following independent binary feature random variables as part of $\mathcal{T}_I$ and $\mathcal{T}_J$: *Rotate left wheel at specified speed*, used at all states and for all actions except turn left; *Rotate right wheel at specified speed*, used at all states and for all actions except turn right; *Navigation ability* that models the robot's localization and plan following capabilities in the absence of motion errors, used at all states and for all actions except stop; *Floor slip*, used for all states and actions. $R(s,a)$ for I and J involves the same binary feature functions as in Bogert and Doshi [2]. For comparison, we consider an approach, labeled as Known R$_{/T}$, that learns the transition function but knows the reward functions of patrollers including how they interact with L acting accordingly. This approach acts as an *upper bound*.

Each robot in our simulations is a **TurtleBot** equipped with a Kinect. ROS's default local motion planner is used for navigation. Each robot localizes itself in a map using the adaptive MCL available in ROS. L is spotted if it is roughly within 6 cells of a patroller and the patroller faces it. We vary the starting locations of the patrollers across runs. We study the impact of mIRL$^*_{/T}$+Int on L's *success rate* in *simulation*. This is the proportion of runs in which L reaches the goal state unspotted by a patroller. Another key metric is the *learned behavior accuracy*, which is the proportion of all states at which the actions prescribed by the inversely learned policy of the patroller coincide with their actual actions. This metric permits focus on the learning in mIRL$^*_{/T}$+Int.

We begin by evaluating the learned behavior accuracy of mIRL$^*_{/T}$+Int as a function of the degree of observability and observing time, in Fig. 1. The degree is the proportion of all (x,y) cells in the state space that are visible to L; its complement gives a measure of the occlusion. Known R$_{/T}$ provides an artificial upper bound. Each data point is the average of 200 simulated runs. Expectedly, the accuracy of mIRL$^*_{/T}$+Int improves with both observability and time. Furthermore, behavior accuracy correlates positively with success rate that reaches up to **60%** for mIRL$^*_{/T}$+Int.

REFERENCES

[1] P. Abbeel and A. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, page 1, 2004.

[2] K. Bogert and P. Doshi. Multirobot IRL with interactions under occlusion. In *AAMAS*, pages 173–180, 2014.

[3] A. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *ICML*, pages 663–670, 2000.

[4] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *IJCAI*, pages 2586–2591, 2007.

[5] B. Ziebart and A. Maas. Maximum entropy inverse reinforcement learning. In *AAAI*, pages 1433–1438, 2008.