Online Inverse Reinforcement Learning Under Occlusion

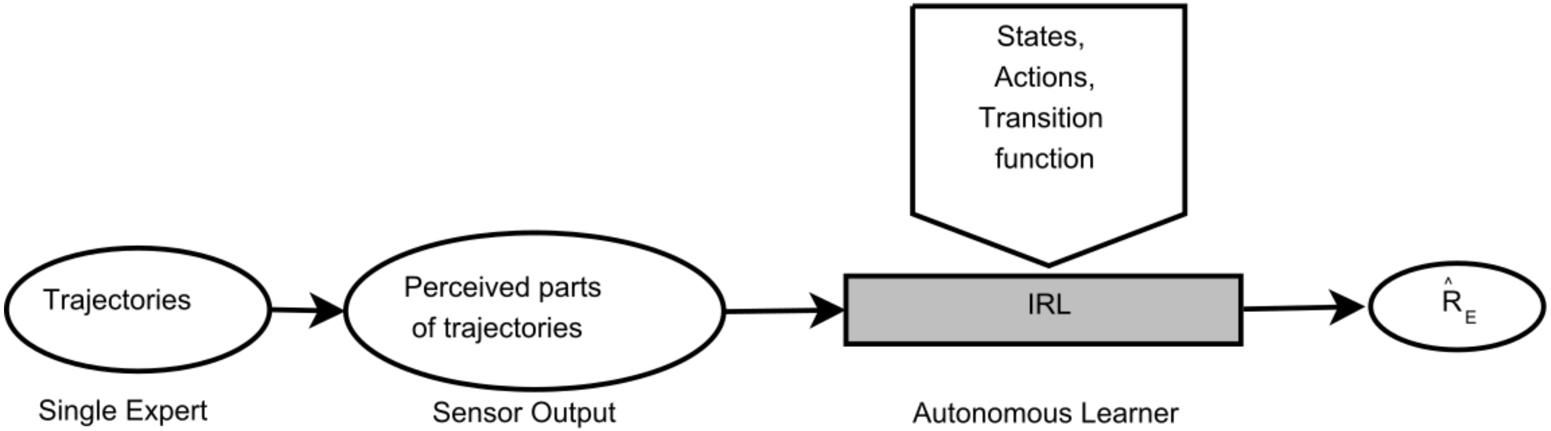
Saurabh Arora and Prashant Doshi
Dept. of Computer Science, University of Georgia
Bikramjit Banerjee
School of Computing Sciences & Computer Engineering
University of Southern Mississippi

Inverse Reinforcement Learning (IRL)

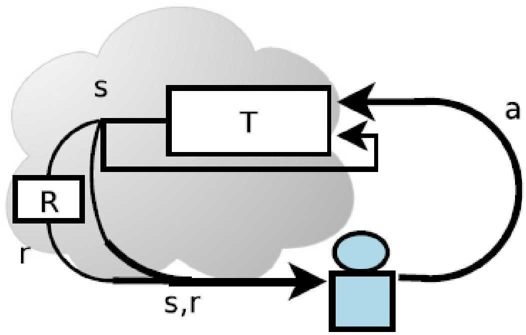
Given: observations of the behavior of an agent engaged in a well-defined task. The observations are in the form of trajectories of state-action pairs

Find: reward function of the agent

Assumption: other parameters of the observed agent are known



Inverse Reinforcement Learning (IRL)



RL

Learner L observes behavior and infers reward function $\hat{R}_E$ of expert E

Linear structure: $\hat{R}_E(s,a) = \theta^T \phi(s,a)$ where θ are weights and ϕ are features

Contributions

- General framework for *incremental IRL* (I2RL)
- Instantiation of I2RL for learning with hidden variables – Latent Max-Entropy I2RL (LME I2RL)
- Formally proved convergence properties (monotonicity and sample complexity bounds)
- Experimental validation of faster convergence of incremental IRL as compared to batch IRL

Incremental IRL (I2RL) Framework

- **Session of IRL:** $\boldsymbol{i}^{\text{th}}$ session of I2RL is a function ξ_i revising the current estimate of the expert's reward function by using as input,
  - the expert's MDP,
  - the curre
  - the rewar

- **I2RL:** Increme inue infinitely or until a stop

A session in online LP-IRL (Jin et al. 10)

$$\xi_i(MDP_{/R_E}, X_i, \hat{R}_E^{i-1})$$

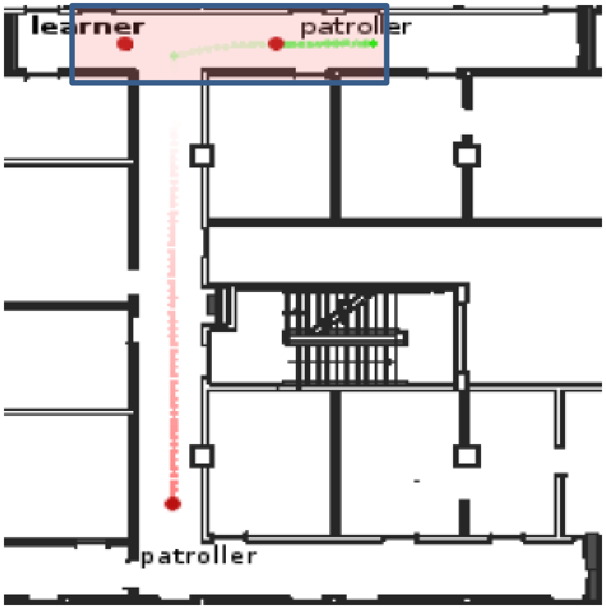
A session in online MaxEnt (Rhinehart&Kitani 17)

$$\xi_i(MDP_{/R_E}, X_i, \theta^{i-1})$$

- **Stopping criteria**: Present two stopping criteria based on the difference in log likelihoods and inverse learning error of the learned reward functions in two successive iterations

IRL with hidden variables

- Learner's observations of the expert may be partially missing due to occlusion
  - Cause could be environment and limits of learner's observational ability
- Presents as missing state-action pairs in the observed trajectory
- *Bogert et al. 16* generalizes maximum-entropy IRL to latent maximum-entropy IRL to allow for hidden variables



Bogert, K., Lin, J. F-S, Doshi, P., and Kulic, D., 2016, May. Expectation-Maximization for Inverse Reinforcement Learning with Hidden Data. In Proceedings of the 17th Conference on Autonomous Agents and Multi Agent Systems (pp. 522-529). International Foundation for Autonomous Agents and Multiagent Systems.

Latent Max-Entropy IRL Formulation

$$\max_{P\in\Delta}\left(-\sum_{X\in\mathbb{X}} P(X;\boldsymbol{\theta})\ log\ P(X;\boldsymbol{\theta})\right)$$

subject to

$$\sum_{X\in\mathbb{X}} P(X;\boldsymbol{\theta}) = 1$$

$$E_{\mathbb{X}}[\phi_k] = \hat{\phi}^{Z|Y}_{\boldsymbol{\theta},k} \qquad \forall k$$

$\hat{\phi}^{Z|Y}_{k}$ - expectation over k^{th} feature computed from observations

$$\hat{\phi}^{Z|Y}_{\boldsymbol{\theta},k} \triangleq \frac{1}{|\mathcal{Y}|}\sum_{Y\in\mathcal{Y}}\sum_{Z\in\mathbb{Z}} P(Z|Y;\boldsymbol{\theta})\sum_{t=1}^{T}\gamma^t\phi_k(\langle s,a\rangle_t)$$

where

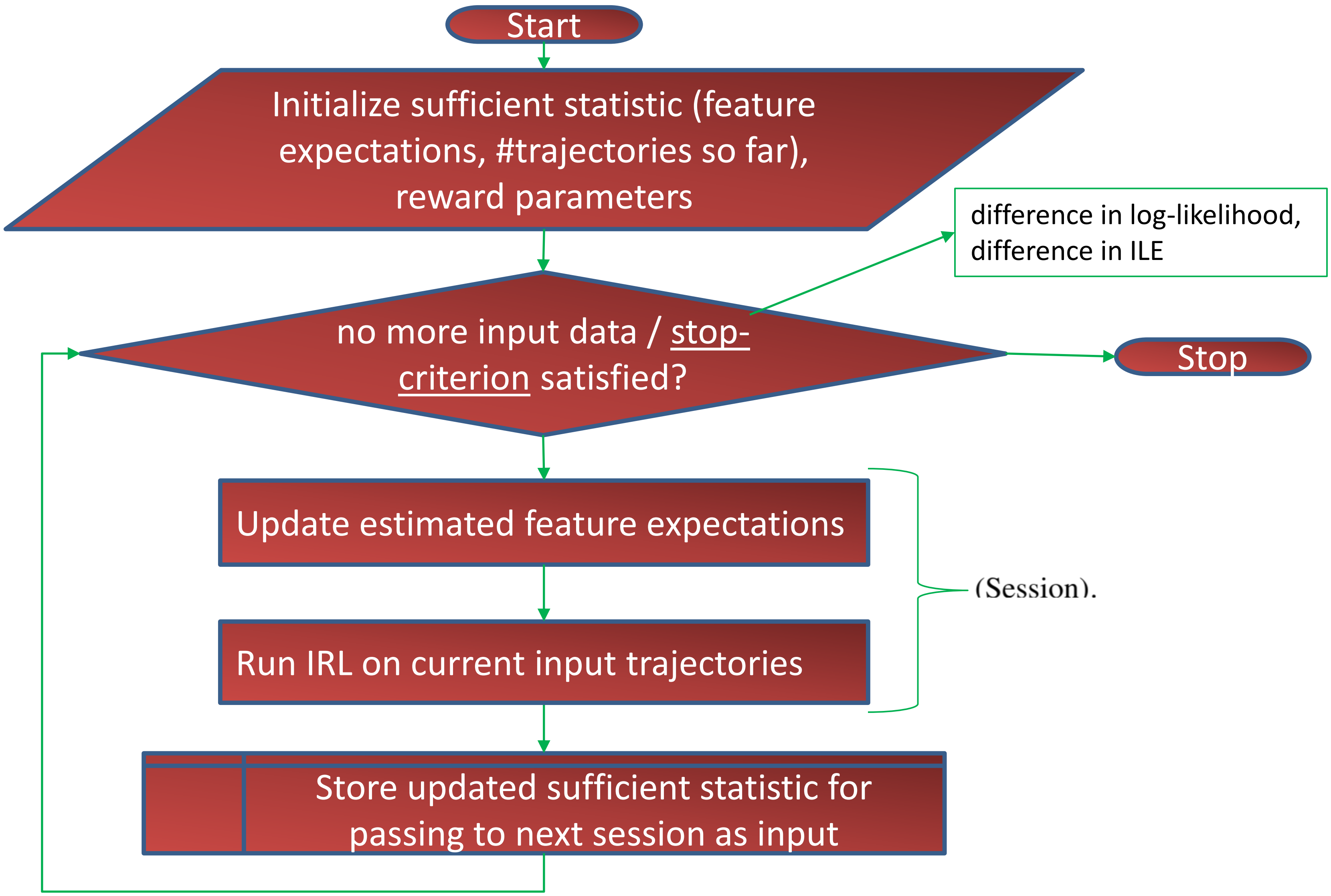
Y - observed part of trajectory;

Z - one of many alternatives for unobserved part;

$X = (Y, Z)$ - one of many ways to complete trajectory;

Learning with hidden variable: *EM formulation of maximum entropy IRL takes expectations over latent variables*

LME I2RL: Online IRL under Occlusion

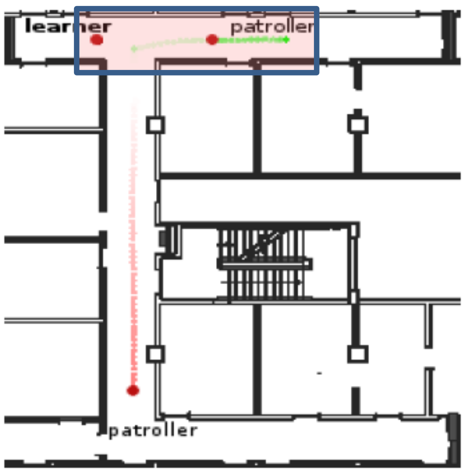


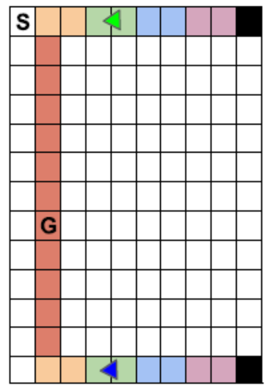
Convergence Properties

- **Estimation error:** When some portion of the demonstration is **hidden** from learner, then the cumulative error in estimating feature values of LME I2RL can be bounded with a probability that depends:
  - linearly on the number of features
  - exponentially on the number of samples for estimating hidden portion
  - allowed error in log likelihood of the learned reward
- **Monotonicity:** After (fully or partially) observing a sufficient amount of trajectories, with each new session the likelihood improves monotonically
- **Convergence:** LME I2RL converges probabilistically in the log-likelihood of learned rewards within an error directly proportional to the number of reward features, error in estimation, and discount factor of MDP

Evaluation: Perimeter Patrol

region of visibility (for physical experiment)





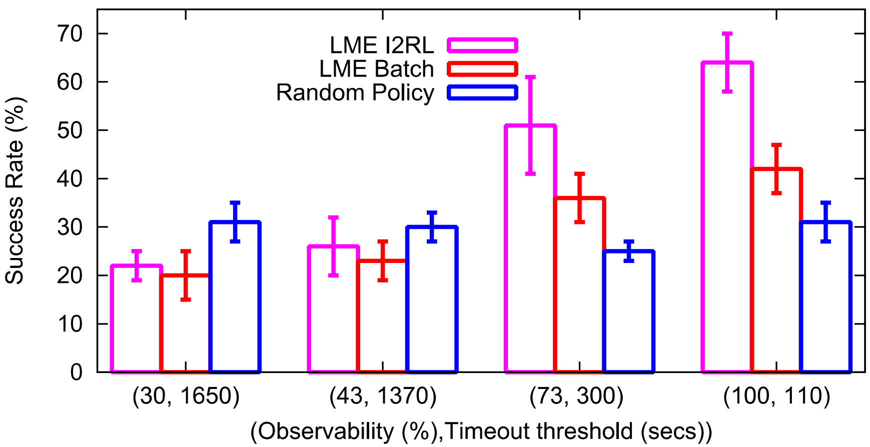
activation regions of our 5 feature functions

visibility in simulation can be varied from 30% to 100%, but it is fixed at 30% for physical experiments

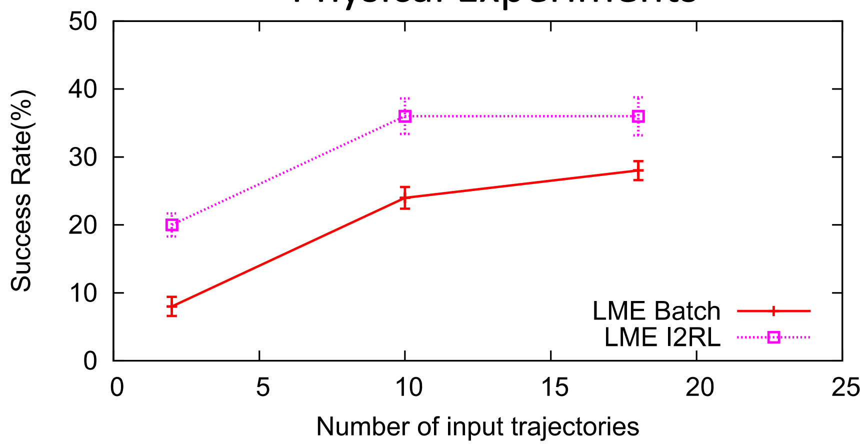
Experimental Results: Success Rate

- In batch IRL, all data is available within one input set, whereas it is given in sessions for incremental
- **Rate of success** is the percentage of attempts for which penetration was successful
- Rate of successful attack for incremental IRL is higher than that for batch IRL

Simulation



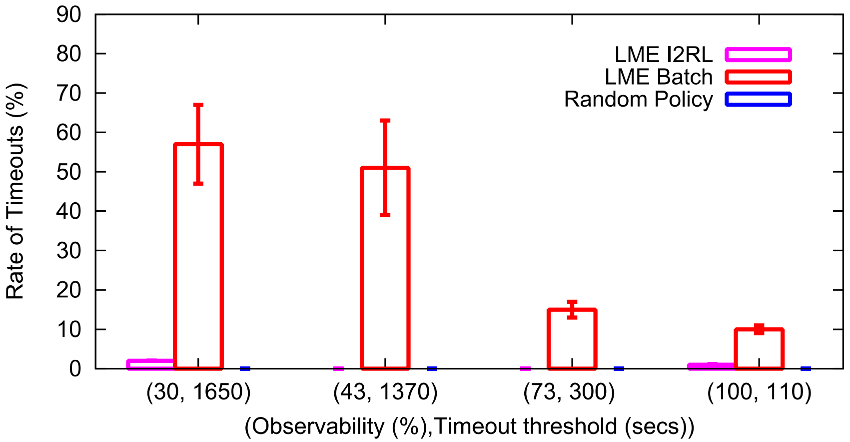
Physical Experiments



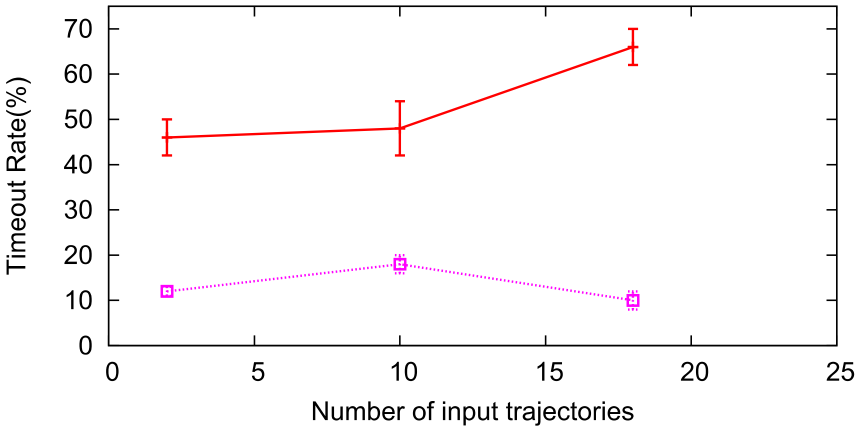
Experimental Results: Timeouts

- **Rate of timeouts** is percentage of runs for which learner failed to move
  - Could not finish learning and planning within time limit
- I2RL cuts down on the time out rate significantly

Simulation



Physical Experiments



Conclusion

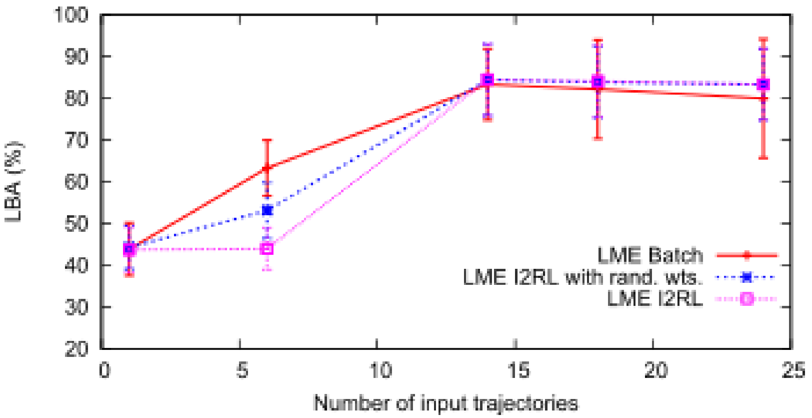
- Defined a generic framework I2RL for online IRL
- Introduced a new method for online IRL with hidden variables – LME I2RL
- Formally proved key convergence properties for LME I2RL
  - monotonic improvement
  - PAC bounds for convergence
- Success rate for I2RL was higher than Batch IRL, primarily because learning finished faster

Thank you (please visit the poster)

Experimental Results: Learning Accuracy

- I2RL is successful in learning the behavior of patrollers
- Learning improves monotonically across sessions

Observability 30%



Observability 70%