Using Active Learning for Inverse Reinforcement Learning

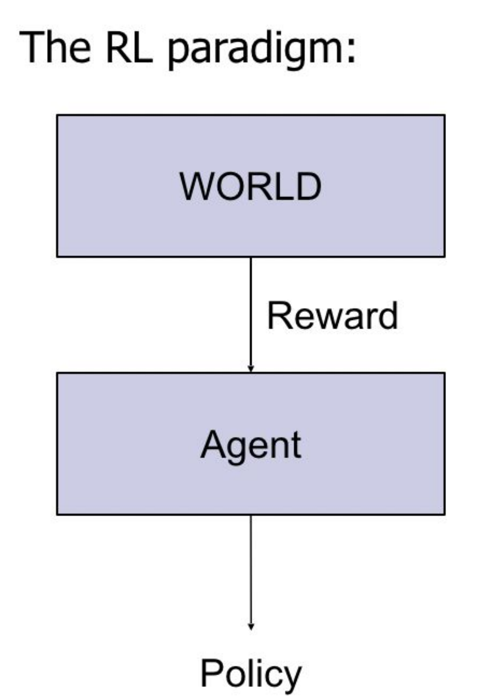
Radhika Nayar, Rohan Dutta, Vivek Krishnamurthy

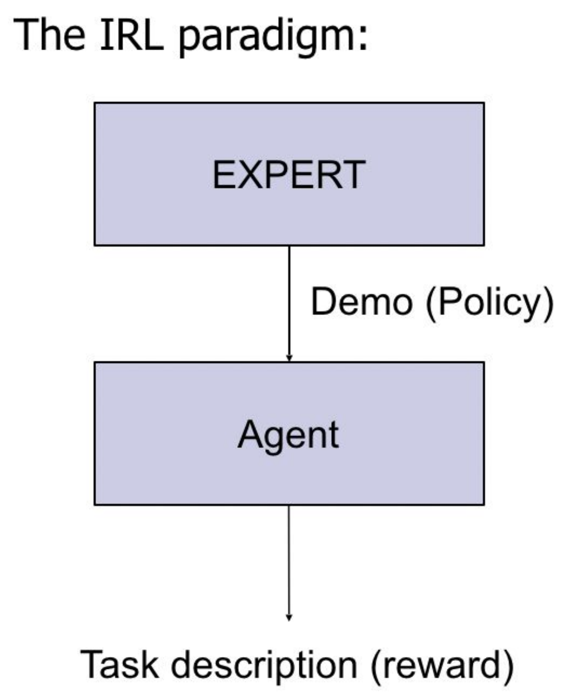
07 Dec, 2019

Questions we are trying to answer from our literature review:

- Why is active learning used in inverse reinforcement learning?
- What kind of research (old/latest) has been done in the field?
- What is the current state-of-the-art in this field?
- What general inferences can be drawn about this field as a whole?
- Where does the field lack progress and why?

Reinforcement Learning (RL) VERSUS Inverse Reinforcement Learning (IRL)



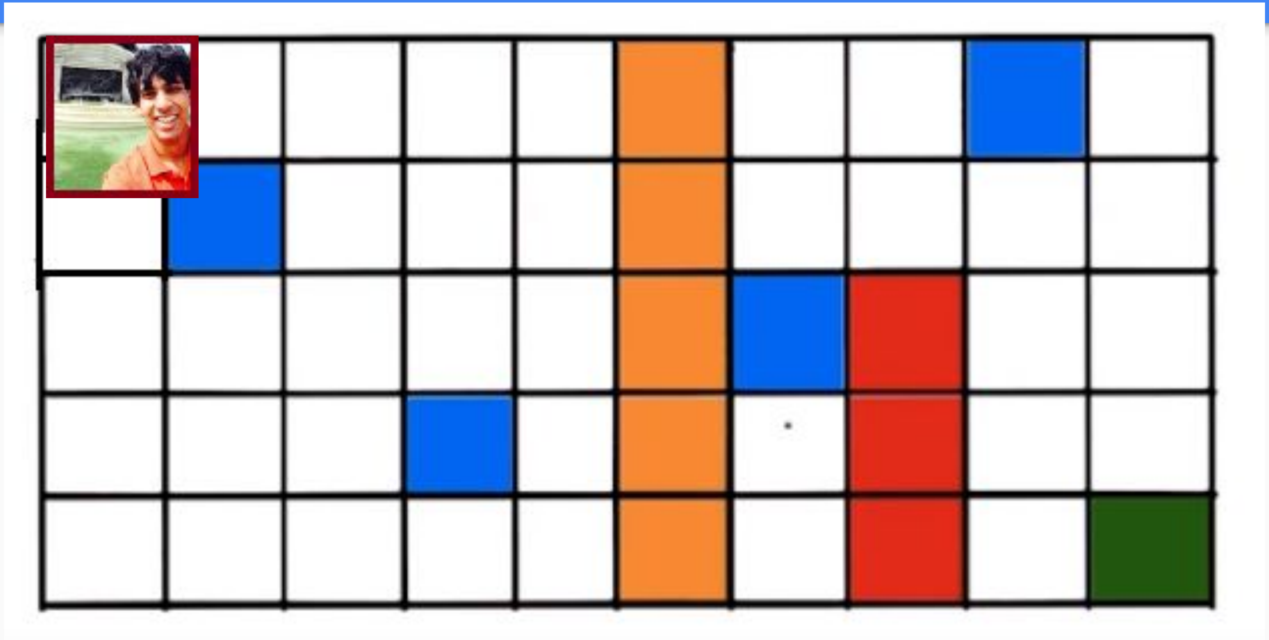


RL: Determine policy that maximizes the total (expected) reward

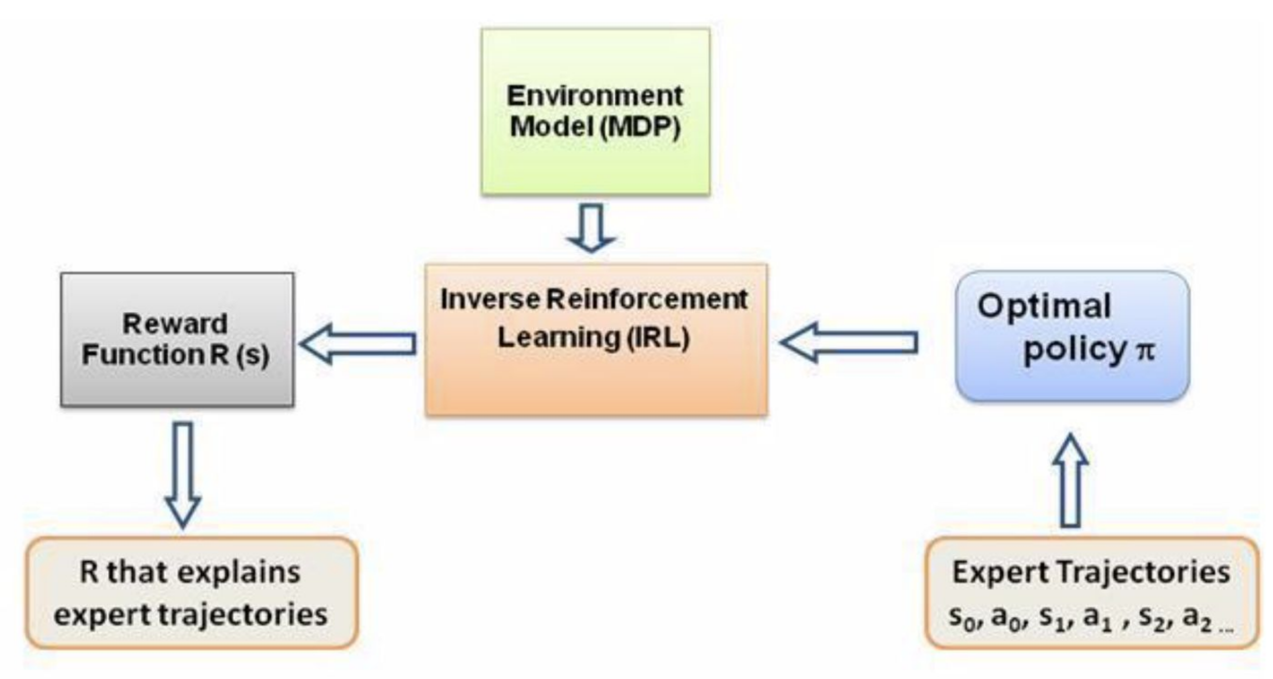
IRL: Estimate the reward function given the samples of policy given by expert

http://videolectures.net/ecmlpkdd09_melo_alreirl/

What is Inverse RL?



Inverse Reinforcement Learning (IRL)


Environment Model (MDP)
Inverse Reinforcement Learning (IRL)
Reward Function R (s)
Optimal policy π
R that explains expert trajectories
Expert Trajectories
$s_0, a_0, s_1, a_1, s_2, a_2$ ...

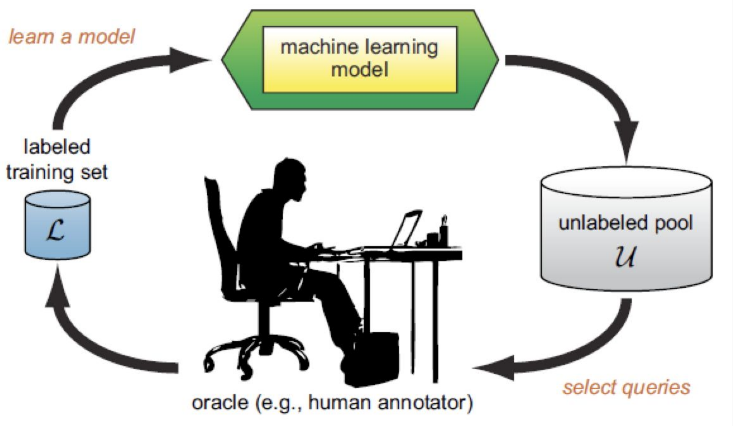
https://www.researchgate.net/figure/Inverse-Reinforcement-learning-Cornell-University-2011_fig3_316786383

- IRL is a form of learning from demonstration
- IRL: finds a reward function R* that explains the expert behaviour

I HAVE NO IDEA
I AM TRYING RANDOM POLICIES
I WILL IMITATE THE EXPERT
I WILL FIND WHAT MOTIVATES THE EXPERT
imgflip.com

Active Learning

- Active learning is a learning algorithm that works by selecting a small but representative set of samples to be labelled by the expert.
- It has two functions:
  - Learn a more accurate classifier and
  - minimize the number of state-action samples needed from the expert.

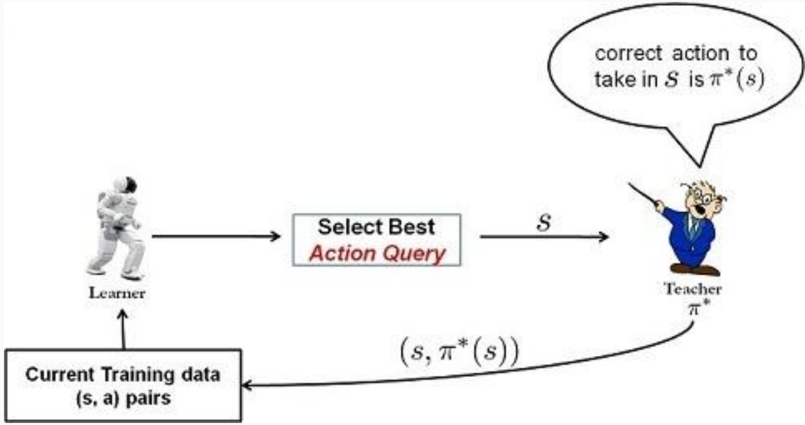


http://videolectures.net/ecmlpkdd09_melo_alreirl/

Objective

Motivation:

- IRL is an ill-defined problem
  - One reward function → multiple optimal policy
  - One optimal policy → multiple reward functions
- Complete demonstrations are often impractical

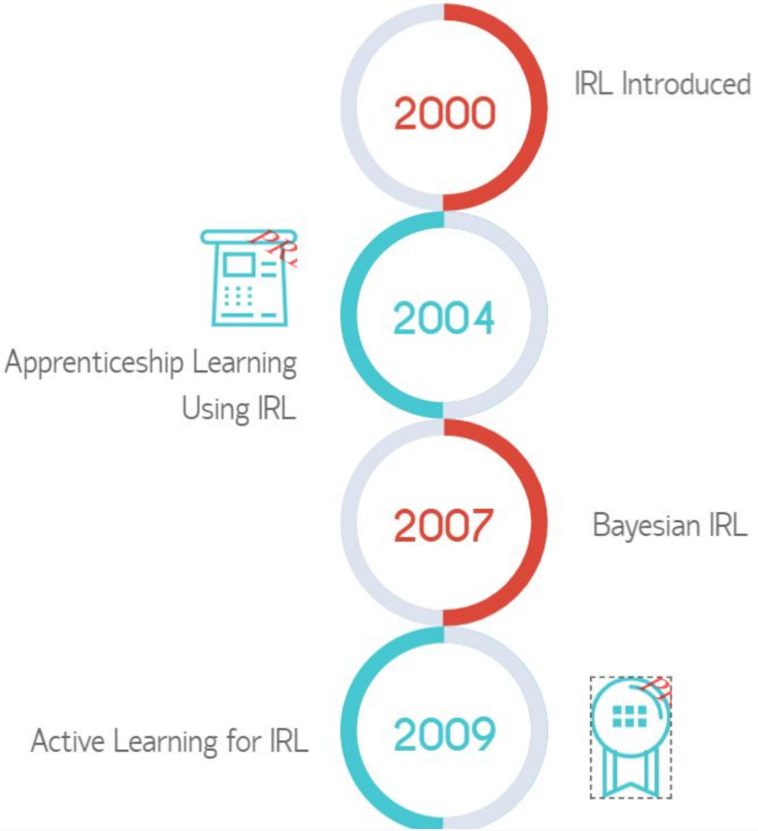


Active Learning + IRL

- Reduces number of demonstrations required from the expert
- Actively query the expert to demonstrate the desired behaviour at the most informative states
- i.e. the learner gains the ability to choose “best” behaviours to be demonstrated by the expert

http://videolectures.net/ecmlpkdd09_melo_alreirl/

Timeline of learnings from demonstration using IRL

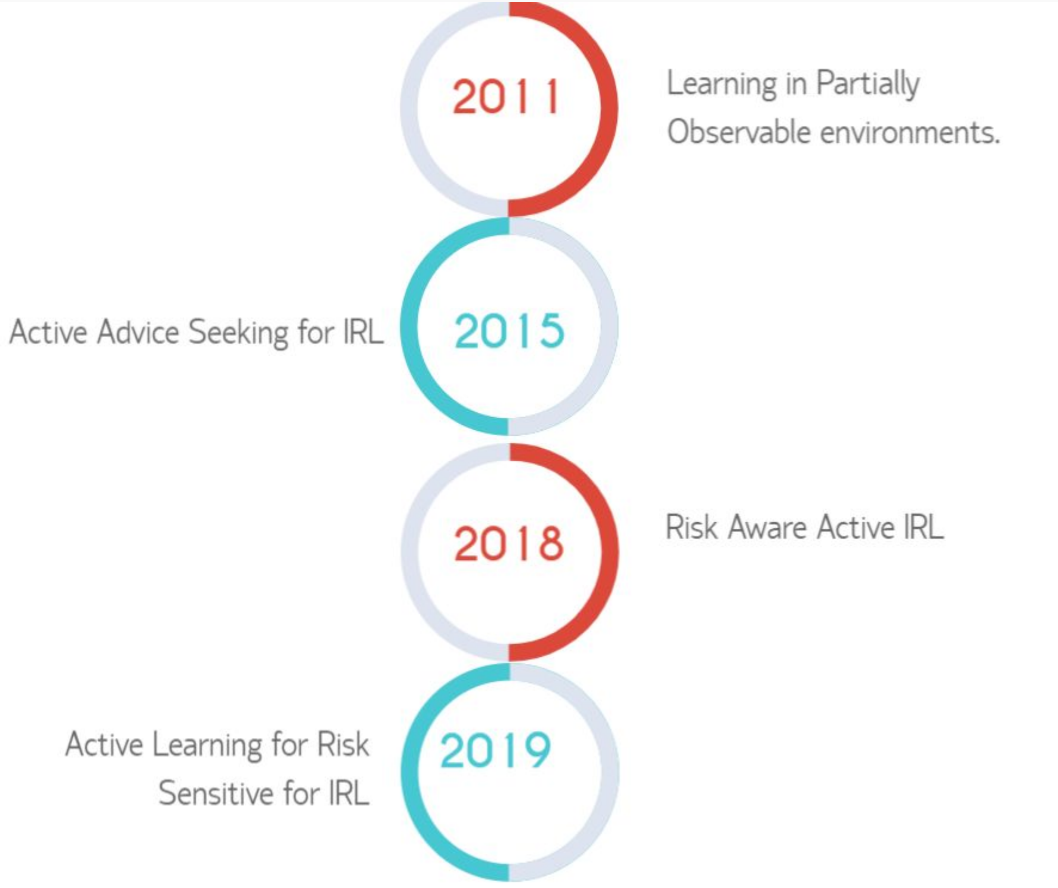


2000 IRL Introduced

2004 Apprenticeship Learning Using IRL

2007 Bayesian IRL

2009 Active Learning for IRL



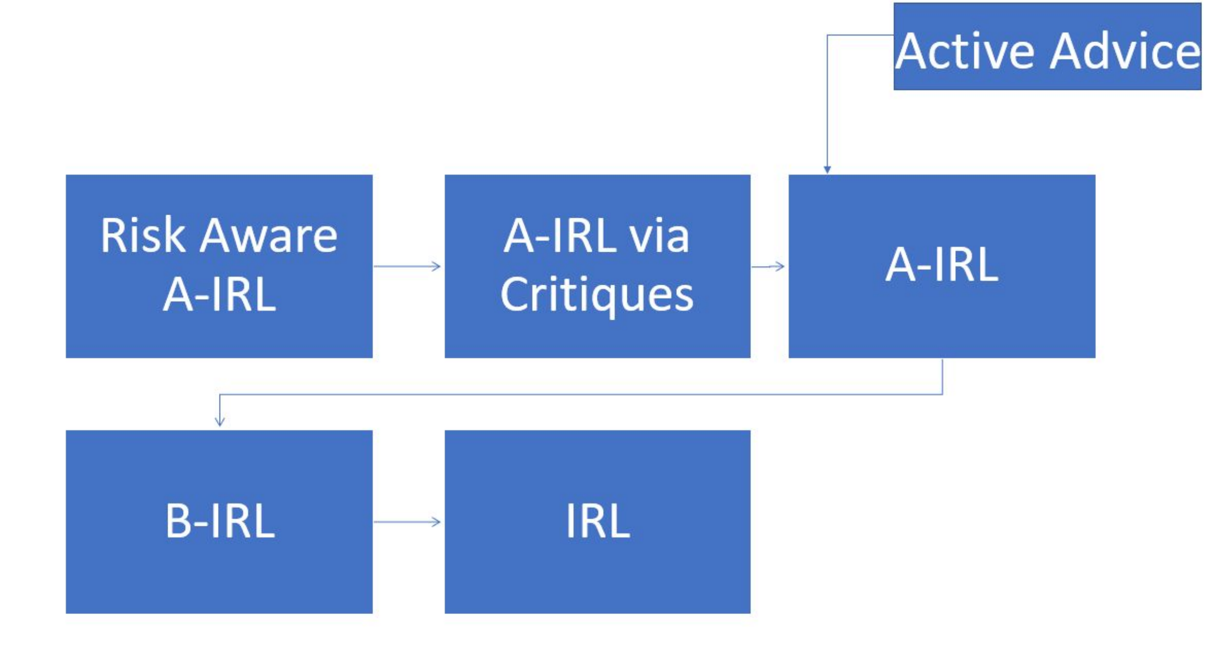
2011 Learning in Partially Observable environments.

2015 Active Advice Seeking for IRL

2018 Risk Aware Active IRL

2019 Active Learning for Risk Sensitive for IRL

Concept Wise Flow



Active Advice

Risk Aware A-IRL

A-IRL via Critiques

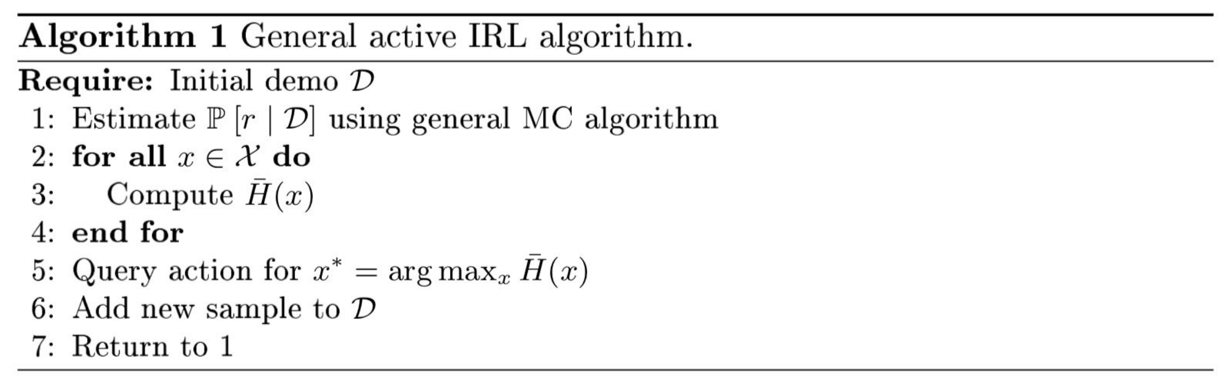
A-IRL

B-IRL

IRL

Seminal Paper: Active Learning + IRL: using Shannon entropy

Select the state to query the expert based on Shannon entropy



Algorithm 1 General active IRL algorithm.

```
Require: Initial demo $\mathcal{D}$
1: Estimate $\mathbb{P}[r \mid \mathcal{D}]$ using general MC algorithm
2: for all $x \in \mathcal{X}$ do
3:     Compute $\bar{H}(x)$
4: end for
5: Query action for $x^* = \arg\max_x \bar{H}(x)$
6: Add new sample to $\mathcal{D}$
7: Return to 1
```

Compute per state average entropy

$$H(x) = {}^{1}/_{|A|} \sum_a H(\mu_{xa})$$

Lopes, F.S. Melo, L. Montesano, Active learning for reward estimation in inverse reinforcement learning, in: European Conference on Machine Learning,ECML/PKDD,Bled,Slovenia,2009.

Active Advice Seeking for IRL

Traditional Active Learning	**Active Advice Seeking**
Constrained to selecting a single example to label	An active advice seeker can get advice over a larger section of the feature space
Active learning is not as expressive.	Advice can be much more expressive than just a single label, example: grouping a set of classes.
Learner queries only a single state	This allows learner to query similar areas of feature space together potentially reducing the number of advice that the learning algorithm requires. Reduces burden of the expert.

Odom P and Natarajan S (2015) Active advice seeking for inverse reinforcement learning. In: *Proceedings of AAAI*, pp. 4186–4187.

Advice Seeking Algorithm

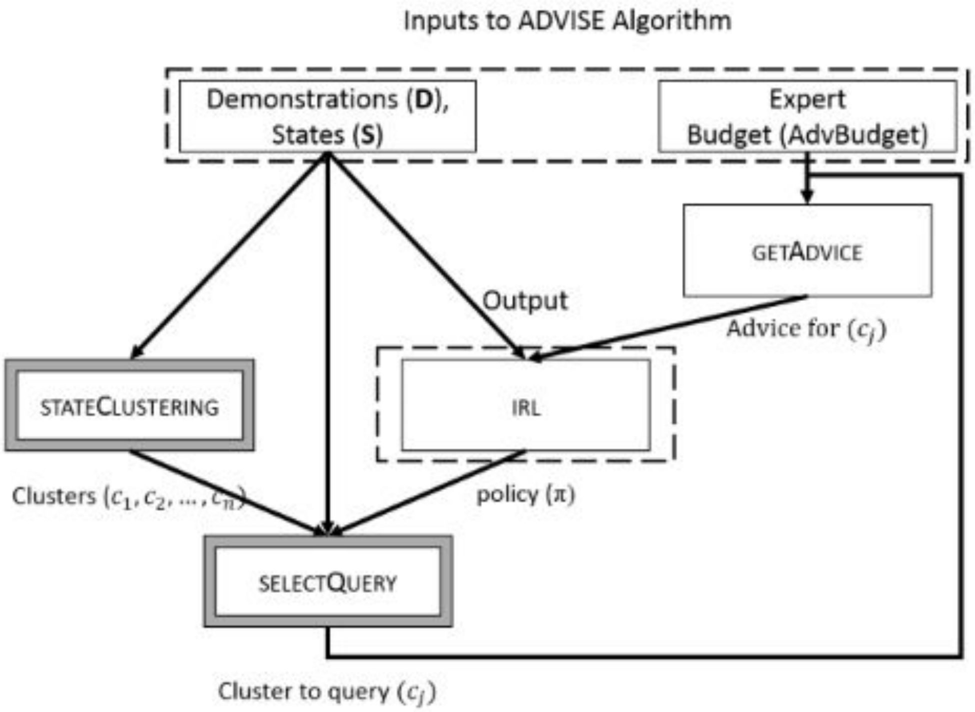

Inputs to ADVISE Algorithm
Demonstrations (**D**), States (**S**)
Expert Budget (AdvBudget)
GETADVICE
Output
Advice for (c_j)
STATECLUSTERING
IRL
Clusters $(c_1, c_2, \dots, c_n)$
policy (π)
SELECTQUERY
Cluster to query (c_j)

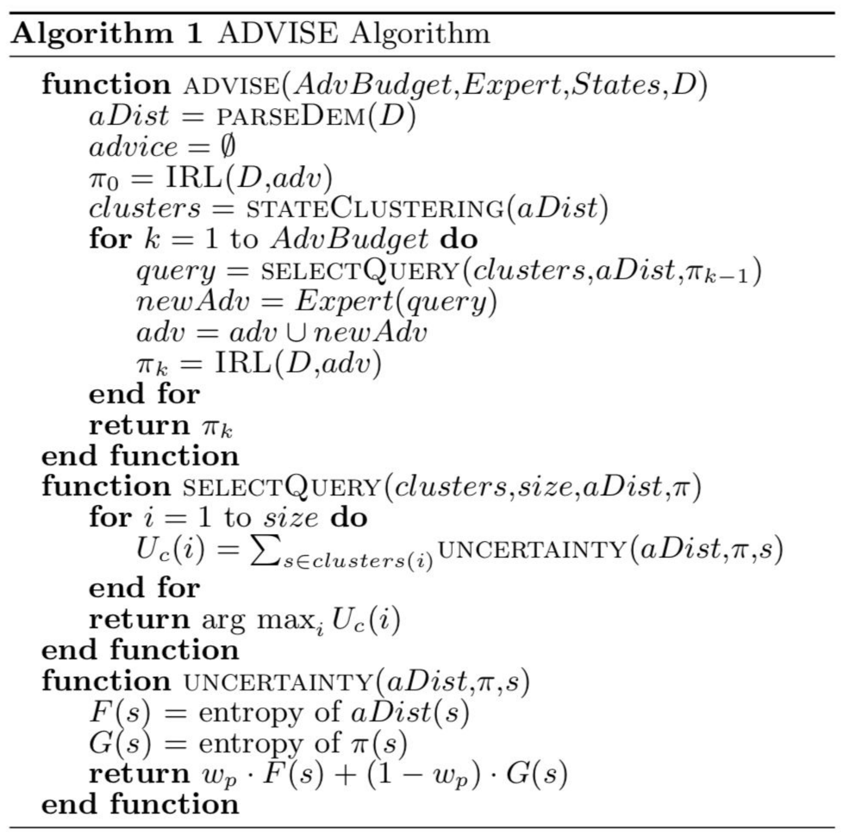
Figure 1: A framework for active advice-seeking.

Odom P and Natarajan S (2015) Active advice seeking for inverse reinforcement learning. In: *Proceedings of AAAI*, pp. 4186–4187.

ADVISE Algorithm (Active aDVIce SEeking): STATECLUSTERING + STATEQUERY

$$U_s(s_i) = w_p F(s_i) + (1 - w_p) G(s_i)$$

$$U_c(\boldsymbol{c}_j) = \frac{\sum_{s_i \in \boldsymbol{c}_j} U_s(s_i)}{|\boldsymbol{c}_j|}$$

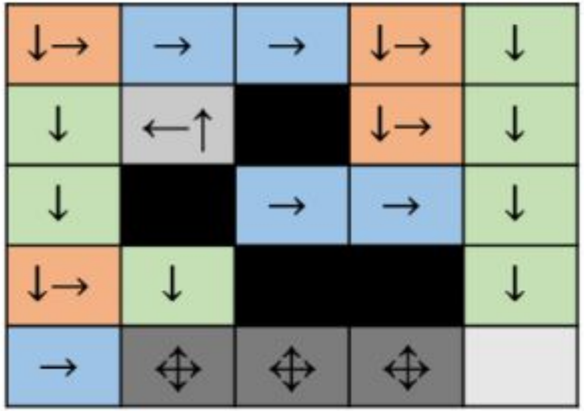


Algorithm 1 ADVISE Algorithm

```
function ADVISE($AdvBudget$,$Expert$,$States$,$D$)
    $aDist$ = PARSEDEM($D$)
    $advice = \emptyset$
    $\pi_0$ = IRL($D$,$adv$)
    $clusters$ = STATECLUSTERING($aDist$)
    for $k = 1$ to $AdvBudget$ do
        $query$ = SELECTQUERY($clusters$,$aDist$,$\pi_{k-1}$)
        $newAdv$ = $Expert$($query$)
        $adv = adv \cup newAdv$
        $\pi_k$ = IRL($D$,$adv$)
    end for
    return $\pi_k$
end function
function SELECTQUERY($clusters$,$size$,$aDist$,$\pi$)
    for $i = 1$ to $size$ do
        $U_c(i) = \sum_{s \in clusters(i)}$ UNCERTAINTY($aDist$,$\pi$,$s$)
    end for
    return $\arg\max_i U_c(i)$
end function
function UNCERTAINTY($aDist$,$\pi$,$s$)
    $F(s)$ = entropy of $aDist(s)$
    $G(s)$ = entropy of $\pi(s)$
    return $w_p \cdot F(s) + (1 - w_p) \cdot G(s)$
end function
```

Odom P and Natarajan S (2015) Active advice seeking for inverse reinforcement learning. In: *Proceedings of AAAI*, pp. 4186–4187.

Wumpus World Example



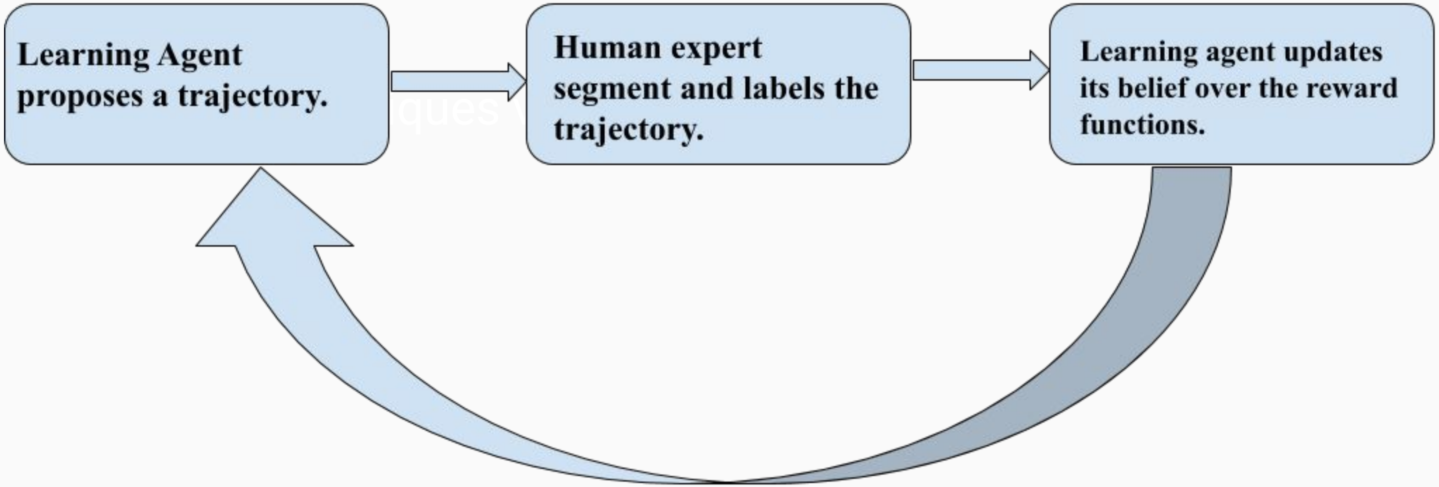
↓→	→	→	↓→	↓
↓	←↑	■	↓→	↓
↓	■	→	→	↓
↓→	↓	■	■	↓
→	✥	✥	✥	

Figure 2: Illustrative example of the ADVISE algorithm for the WumpusWorld domain. The arrows show the actions demonstrated by the human expert. The colors show an ideal clustering of the states. Ideally the algorithm will select the dark grey cluster early to correct the sub-optimal actions.

Odom P and Natarajan S (2015) Active advice seeking for inverse reinforcement learning. In: *Proceedings of AAAI*, pp. 4186–4187.

Active Learning from Critiques via Bayesian IRL

- Learner proposes a new trajectory instead of the expert giving the learner sample trajectories.
- The role of the expert is to critique the query.
  - i.e. the expert splits the trajectory into good and bad segments
- Learner updates its belief state over the reward functions.



Learning Agent proposes a trajectory.

Human expert segment and labels the trajectory.

Learning agent updates its belief over the reward functions.

Yuchen Cui and Scott Niekum .2017.Active Learning from Critiques via Bayesian Inverse Reinforcement Learning. In Robotics: Science and Systems Workshop on Mathematical Models, Algorithms, and Human-Robot Interaction

Active Learning from Critiques via Bayesian IRL

Learner proposes a trajectory based on the information gain of a specific state-action pair using the weighted Kullback-Leibler (KL) divergence.

Below is the equation for computing KL-divergence for two distributions p (updated state) and q (original state):

$$D_{KL}(p||q) = \sum_{c=1}^{C} p(c) \log \frac{p(c)}{q(c)}$$

Yuchen Cui and Scott Niekum .2017.Active Learning from Critiques via Bayesian Inverse Reinforcement Learning. In Robotics: Science and Systems Workshop on Mathematical Models, Algorithms, and Human-Robot Interaction.

Risk-Aware Active IRL

- Previous algorithms use some sort of statistical formula to measure the gain in information.

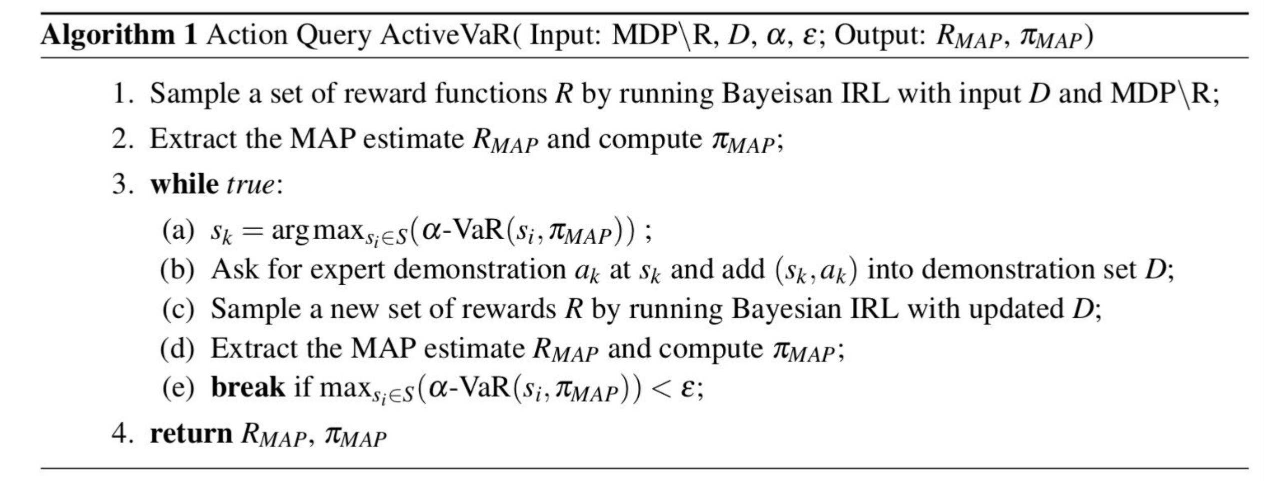
- This algorithm seeks to optimize the policy learnt by the learner, by focusing the queries on areas of the state space with potentially large generalization error.

- It also the first to provide a performance-based stopping criterion that allows a robot to know when it has received enough demonstrations to safely perform a task.

Brown, D. S.; Cui, Y.; and Niekum, S. 2018. Risk-aware active inverse reinforcement learning. In *Proceedings of the 2nd Annual Conference on Robot Learning (CoRL)*.

State-of-the-art: Risk-Aware Active IRL

- Learner generates active queries based on the metric alpha-VaR
- alpha-VaR can be thought of as the risk associated with the state
- Goal is to select the state with highest alpha-VaR (highest risk) to query the expert



Algorithm 1 Action Query ActiveVaR(Input: MDP\R, D, α, ε; Output: R_{MAP}, π_{MAP})

```
1. Sample a set of reward functions $R$ by running Bayeisan IRL with input $D$ and MDP\R;
2. Extract the MAP estimate $R_{MAP}$ and compute $\pi_{MAP}$;
3. while true:
    (a) $s_k = \arg\max_{s_i \in S}(\alpha\text{-VaR}(s_i, \pi_{MAP}))$ ;
    (b) Ask for expert demonstration $a_k$ at $s_k$ and add $(s_k, a_k)$ into demonstration set $D$;
    (c) Sample a new set of rewards $R$ by running Bayesian IRL with updated $D$;
    (d) Extract the MAP estimate $R_{MAP}$ and compute $\pi_{MAP}$;
    (e) break if $\max_{s_i \in S}(\alpha\text{-VaR}(s_i, \pi_{MAP})) < \varepsilon$;
4. return $R_{MAP}$, $\pi_{MAP}$
```

where MAP: Maximum A Posteriori (In this context, our estimate for the most likely reward function given our current knowledge)
where ε is user defined threshold.

Brown, D. S.; Cui, Y.; and Niekum, S. 2018. Risk-aware active inverse reinforcement learning. In *Proceedings of the 2nd Annual Conference on Robot Learning (CoRL)*.

Risk-Aware Active IRL

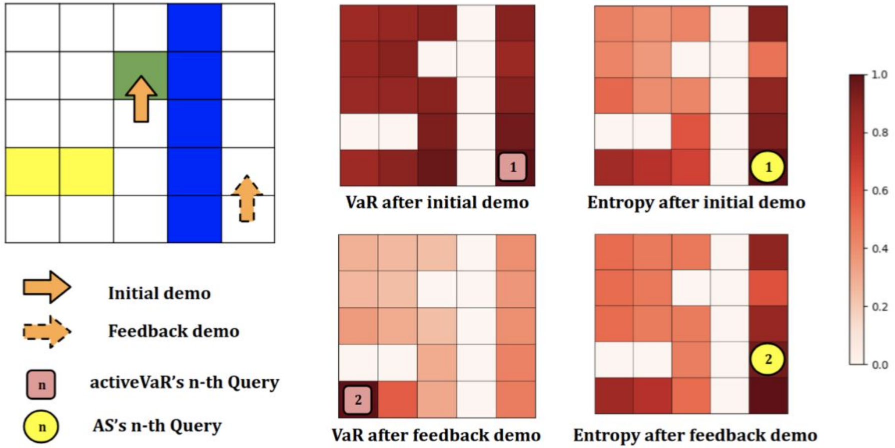


Figure 1: Comparison of active action queries based on performance loss risk or action entropy. The example gridworld has four different unknown features denoted by the yellow, green, white, and blue colors of the cells. White states are legal initial states. AS is the active learning algorithm proposed by Lopes et al. [5] and activeVaR is our proposed active action query algorithm. The first two active queries proposed by each algorithm are annotated on the heatmaps of VaR and entropy values after each iteration. For heatmaps, all values are normalized from 0 to 1.

Brown, D. S.; Cui, Y.; and Niekum, S. 2018. Risk-aware active inverse reinforcement learning. In *Proceedings of the 2nd Annual Conference on Robot Learning (CoRL)*.

Conclusion

- All the algorithms are built upon Bayesian IRL in which the learner estimates the posterior probability over possible reward functions, given the demonstration.
- All papers define some metric to “select” states based on the criterion that gives a measure of maximum information gained by querying a particular state.
- The risk aware strategy is the only algorithm that is concerned with how good the actual policy executed by the robot is and how well it generalizes; the other algorithms are only concerned about some statistical metric.

Where does the field lack progress and why?

- An inherent drawback of IRL is that it is ill posed: for a single policy there can be multiple rewards and for a single reward there can be multiple policies. This means that we have a large space to optimize over.

- Main drawback with all active learning algorithms is that they assume that the learner has complete access to the state space and can query the expert for a demonstration on any of the states.

- AI research has been primarily focused in areas of supervised learning problems where we have a well defined input output (such as face and recognition, video captioning and speech processing). Active learning is ill suited for these problems compared to supervised learning algorithms.

- In other areas such as image generation, active learning loses out to unsupervised learning methods such as (Generative Adversarial) GAN’s. It is very hard to define these problems in terms of Active learning.

- In the current literature, active learning has been used in grid world simulations, simulated 2D navigation, and robot table setting tasks. These domains tend to be very narrow and confined. However, in real world scenarios, the number of possible states increase greatly.

Future Improvements

- Since all of the algorithms are based on using a Bayesian prior, when the prior is decorrelated from the reward in different states, we do not expect active learning to bring a significant advantage. It would require a fundamental change in the way we approach reward estimation.

- More research required on deciding the “best” criterion for deciding which state-action pair query would be most beneficial.

- More research required on understanding relational techniques to increase the expressiveness of query to the expert.

- Most of the Active Learning use cases tend to be simulations on a computer. Some of the active Learning applications in real life scenarios are drone navigation and placing objects on a table . More research can be done on sophisticated real life demonstrations.

Hasta La Vista