https://thegradient.pub/learning-from-humans-what-is-inverse-reinforcement-learning/

As Ng and Russell put it, “**the reward function, rather than the policy, is the most succinct, robust, and transferable definition of the task**,”

 The main problem when converting a complex task into a simple reward function is that **a given policy may be optimal for many different reward functions**. That is, even though we have the actions from an expert, there exist many different reward functions that the expert might be attempting to maximize

Some of these functions are just silly: for example, all policies are optimal for the reward function that is zero everywhere, so this reward function is always a possible solution to the IRL problem. But for our purposes, we want a reward function that captures meaningful information about the task and is able to differentiate clearly between desired and undesired policies.

To solve this, Ng and Russell formulate inverse reinforcement learning as an optimization problem. We want to choose a reward function for which the given expert policy is optimal. But given this constraint, we also want to choose a reward function that additionally maximizes certain important properties.

Well explains algorithm

https://thinkingwires.com/posts/2018-02-13-irl-tutorial-1.html

Explains mathematics

<https://www.youtube.com/watch?v=J2blDuU3X1I&list=PLkFD6_40KJIwTmSbCv9OVJB3YaO4sFwkX&index=14>

jn-survey

two reward functions similar for

the most part but di\_ering for some state-action pairs may produce considerably

di\_erent policies (behaviors). To make the evaluation targeted, a comparison of

the behavior generated from the learned reward function with the true behavior

of expert is more appropriate.

Good generalised template of IRL is given

Challenges

Curse of Dimensionality

In example like Robots, ‘Continuous space’. Also if discretized leads to ‘(almost) infinite space’. That leads to Time complexity of algorithm.

Sample complexity: increase in required sample trajectories as problem complexity increases.

Perturbations: noisy demonstrations

Incomplete and Imperfect observations/ IRL with occlusion

POMDP paper

Draw inferences about unobserved state-action pairs

All methods develop on MAXENTIRL

(already read)Un/ripe fruit picking

Incomplete Model/missing features

Convergence analysis

Direct vs Indirect Learning warrants more attention(no detail info)

Nonlinear reward function

Multi-task/Multi-Reward

Non stationary rewards: reward function that changes with time ()

Multi-expert Interaction.

Algo for IRL

Policy and values fn definition given

Degeneracy problem (eg. 0 and same rewards) & soln (eg. increasing margin,)

Predicting driving behaviour

-modifies MaxEnt equation(clarify from video)

- uses DPM as core strategy (Dirchlet)

Apprenticeship Learning with Multiple intentions:

- e.g. safe driver vs Ambulance

- trajectories are not always labelled

- this algo. clusters (near optimal) trajectories which have similar reward function

-  **θe** (not known) -> (assume) **θa**

- summarizing IRL approach

- Boltzmann exploration

2 algos

1. Linear Programming Apprenticeship Learning(LPAL)

2. MLIRL

Doubt whether it converge for infinite horizon (hence, sample)

Defn of MI-IRL

EM

Application: climate control, intention(normal/threatening) surveillance

Gradient based minimization

Expert share the same reward fn but their policies may be different

Reference to papers with same topic

GitHub repo implemented all IRL algo

<https://github.com/yrlu/irl-imitation>

Explanation and GitHub code: (maths can be taken)

<http://178.79.149.207/posts/maxent.html>

code: <https://github.com/yrlu/irl-imitation>

Multiple intention code

<https://github.com/jmacglashan/burlap/blob/master/src/main/java/burlap/behavior/singleagent/learnfromdemo/mlirl/MultipleIntentionsMLIRL.java>

Heterogeneous Multi-agent Dissertation:

Lit review: def RL, model-based model-free, Q-learning, deep learning+DCN, backpropagation, CNN, DRL, DQN, MARL, DRL in traffic control

Design: traffic ight control problem: state representation(0-1 matrices)

State problem with Reward on page 97

**Ideas**

IRL might be seen as reward shaping utility, instead of just reward extraction. E.g. see the reward maps recovered yet. It shows how the path is gradually coloured, instead of just simply showing the goal state in Yellow.