Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 22

Solution by : Amirhossein Asadi

Q1

The Maximum Margin Principle in IRL aims to:
A) Minimize the number of expert demonstrations required.
B) Ensure the expert policy's expected reward exceeds all others by a margin m.
C) Reduce the dimensionality of the feature space.
D) Approximate the Q-function using linear regression.

Correct Answer: B) Ensure the expert policy's expected reward exceeds all others by a margin m.

The Maximum Margin Principle in Inverse Reinforcement Learning (IRL) aims to maximize the margin between the expert's expected reward and that of any other policy. This principle helps ensure that the expert's policy is distinguished from other policies by a clear margin, which encourages learning a policy that is close to the expert's behavior. This approach is often used to address the problem of reward specification in IRL, where the goal is to learn a reward function that captures the expert's intentions and behavior while distinguishing it from other possible policies.

Q3

When an IRL algorithm for a navigation robot proposes multiple different reward functions that all equally justify the expert's behavior, this demonstrates which core challenge?
A) Overfitting to expert data.
B) Under-specification problem.
C) Computational instability in optimization.
D) Dependency on handcrafted features.

Correct Answer: B) Under-specification problem.

The under-specification problem in Inverse Reinforcement Learning (IRL) occurs when multiple reward functions can explain the same expert behavior equally well, leading to ambiguity in selecting the correct reward function. In this case, the IRL algorithm is unable to uniquely identify a single reward function that represents the expert's true intentions, because the available expert data does not provide sufficient constraints to narrow down the possible reward functions. This challenge highlights the importance of incorporating additional information, such as priors or further constraints, to resolve the ambiguity and ensure a more robust reward specification.

Q3

In an IRL experiment, the algorithm copies expert behavior through direct imitation rather than learning the true reward function. What likely causes this?
A) Reward-dynamics entanglement
B) Defects in expert data
C) Inappropriate feature selection
D) Improper learning rate tuning

Correct Answer: A) Reward-dynamics entanglement

In Inverse Reinforcement Learning (IRL), reward-dynamics entanglement occurs when the algorithm confuses the learned dynamics of the environment with the reward structure. This happens because the algorithm may directly imitate the expert's behavior without properly distinguishing between the reward and the system dynamics. Instead of learning a true reward function, it mimics the actions of the expert as a behavioral policy, which leads to the copying of the expert's actions without understanding the underlying reward structure. To avoid this issue, careful separation of the reward model and the system dynamics is needed during training.

Q4

A self-driving car company uses IRL to learn reward functions from human drivers. Their algorithm:
- Uses trajectory ranking with neural network rewards.
- Assumes optimality at each timestep.
- Observes that the learned policy often makes abrupt lane changes.

How could you modify the IRL framework to prevent this behavior?
A) Remove entropy regularization
B) Increase the batch size during training
C) Switch to linear reward parameterization
D) Add a kinematic feasibility penalty to r

Correct Answer: D) Add a kinematic feasibility penalty to r

The observed abrupt lane changes are likely a result of the learned policy overfitting to the trajectory ranking model without considering the kinematic constraints of the car. To prevent such behavior, adding a kinematic feasibility penalty to the reward function r would encourage the policy to respect the physical limitations of the vehicle, such as safe steering angles and velocity limits. This penalty ensures that the policy learns to make more gradual and feasible lane changes, rather than unrealistic or abrupt maneuvers that may not be physically possible or safe.

Q5

A kitchen robot learns from human demonstrations but spills liquids when cup sizes change. What IRL limitation caused this?

Answer: Insufficient generalization to unseen variations in the environment.

The issue arises because the robot has likely overfitted to the specific conditions of the demonstrations (e.g., fixed cup size) without learning a generalizable reward function. In this case, the robot did not learn to account for the relationship between cup size and pouring behavior. This is a common limitation in Inverse Reinforcement Learning (IRL), where the model fails to generalize well to variations in the environment that were not present in the expert demonstrations. To resolve this, techniques like domain randomization or adding more diverse demonstrations could help the robot learn to adapt to different cup sizes without spilling.

Q6

An IRL treatment recommender prescribes extreme doses after ICU training. Best solution:
A) More diverse training data
B) Adding safety constraints to reward function
C) Reducing state space dimensionality
D) Using deterministic policies

Correct Answer: B) Adding safety constraints to reward function

The issue of extreme dosing in the treatment recommender can be attributed to a lack of safety considerations during the learning process. In this case, adding safety constraints to the reward function would prevent the model from prescribing unsafe or extreme treatments by incorporating penalties for actions that lead to unsafe outcomes. This ensures that the learned policy not only maximizes the expected reward but also adheres to safety protocols, preventing risky or harmful decisions from being made, especially in critical environments such as the ICU.