Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 21

Solution by : Amirhossein Asadi

Q1

Which method avoids distributional shift when combining offline expert and suboptimal data?
A) BC + Fine-tuning
B) ValueDICE
C) DAgger
D) Random Sampling

Correct Answer: C) DAgger

DAgger is a method designed to mitigate the issue of distributional shift when combining offline expert data with suboptimal data. In typical imitation learning scenarios, policies trained on suboptimal data can cause a mismatch between the expert's behavior and the learned policy. DAgger overcomes this by iteratively collecting new data using the current policy, which is then aggregated with the previous dataset. This ensures that the policy is exposed to states and actions it may encounter in future iterations, helping prevent distributional shift. The key idea is to continually refine the dataset by blending the expert data with data generated by the current policy, thereby minimizing the discrepancies between the expert's distribution and the policy's distribution.

Q2

When applying DAgger to a nonlinear control system, how does the adjoint state from Pontryagin's principle influence the aggregation strategy?
A) It determines which expert actions to trust during data aggregation
B) It provides sensitivity gradients for correcting compounding errors
C) It replaces the need for expert supervision
D) DAgger is fundamentally incompatible with PMP

Correct Answer: B) It provides sensitivity gradients for correcting compounding errors

When applying DAgger to nonlinear control systems, the adjoint state from Pontryagin's Maximum Principle (PMP) provides sensitivity gradients that are essential for correcting compounding errors in the learning process. These gradients help adjust the policy by understanding how small deviations in the control action affect the system's trajectory, especially when integrating expert actions over time. By incorporating this information into the aggregation strategy, DAgger ensures more accurate data collection and minimizes the errors that accumulate as the policy improves. This approach allows DAgger to handle complex control tasks with greater precision and stability.

Q3

In your opinion which method works best for transferring an imitation policy from robot A (dynamics A) to robot B (dynamics B)?
A) Fine-tuning the policy with limited data from robot B
B) Using domain randomization on dynamics A
C) Learning a mapping between state spaces of A and B
D) Combining inverse dynamics with meta-learning

Correct Answer: C) Learning a mapping between state spaces of A and B

Transferring an imitation policy from robot A to robot B can be a challenging task due to the differences in their dynamics. One effective method is learning a mapping between the state spaces of robot A and robot B. This approach involves finding a correspondence between the states and actions in the two robots' environments, allowing the policy learned on robot A to be applied to robot B. By aligning the state representations, the policy can be transferred with minimal modifications, ensuring that robot B can perform similar tasks as robot A despite the differences in their dynamics.

Q4

In the context of IL, particularly when utilizing Gaussian Mixture Models (GMMs) for behavior cloning, what does the term 'mean-seeking behavior' refer to?
A) The agent's policy tends to place the mean at the center of the data, which can lead to suboptimal performance.
B) The agent's policy overfits to the most frequent actions in the expert's demonstrations, neglecting less frequent but potentially crucial actions.
C) The agent's policy fails to generalize beyond the expert's demonstrations, resulting in poor performance on unseen states.
D) The agent's policy exhibits high variance, capturing the full spectrum of the expert's behavior without bias.

Correct Answer: B) The agent's policy overfits to the most frequent actions in the expert's demonstrations, neglecting less frequent but potentially crucial actions.

The term 'mean-seeking behavior' in the context of Gaussian Mixture Models (GMMs) for behavior cloning refers to the agent's tendency to focus on the most frequent actions in the expert's demonstrations. GMMs are typically used to model the distribution of expert behavior, and the agent learns to replicate these behaviors. However, this can lead to the agent's policy overfitting to the most common actions, potentially neglecting less frequent but still crucial actions. This bias towards the center of the data (or the most frequent actions) may result in suboptimal performance, as the agent may fail to account for the full range of actions required to handle all situations in the environment.

Q5

Under what conditions does Behavioral Cloning become mathematically equivalent to offline RL with a specific reward function?
A) When using square loss and deterministic experts
B) When data coverage is sufficient for all states
C) When the reward is the log-likelihood of expert actions
D) Never – they're fundamentally different

Correct Answer: C) When the reward is the log-likelihood of expert actions

Behavioral Cloning (BC) becomes mathematically equivalent to offline Reinforcement Learning (RL) when the reward function is set to the log-likelihood of expert actions. In this case, the agent's objective is to maximize the likelihood of taking actions that the expert would take, which is the same as maximizing the cumulative reward based on expert demonstrations. This can be seen as a specific case of offline RL where the reward function is tied directly to the expert's behavior, and the agent learns by imitating the expert's actions as if they were the rewards guiding its behavior. This equivalence holds when the data coverage is sufficient to represent the full state space the agent will encounter.

Q6

When expert demonstrations are noisy, why does increasing model capacity often decrease imitation performance?
A) The policy memorizes outliers instead of learning robust features
B) Regularization terms dominate the loss function
C) Gradient descent becomes unstable
D) This only happens with recurrent policies

Correct Answer: A) The policy memorizes outliers instead of learning robust features

When expert demonstrations are noisy, increasing the model capacity (e.g., by adding more parameters or layers) often leads to worse performance in imitation learning. This happens because a model with higher capacity has more flexibility, and it tends to overfit to the noisy or outlier data, rather than learning generalizable and robust features from the expert demonstrations. Instead of focusing on the essential patterns in the data, the model starts to memorize the specific noisy examples, which degrades its ability to generalize to new or unseen situations. To mitigate this, techniques such as regularization or using simpler models may help to avoid overfitting and improve performance.

Q7

In robotic manipulation, your IL policy exactly replicates expert finger-twitching motions (biomechanical noise). Why does adding more expert data worsen this behavior?
A) The policy interprets noise as intentional due to maximum likelihood.
B) High-capacity networks fit noise before learning true features.
C) L2 action regularization amplifies minor fluctuation.
D) This is impossible with proper data augmentation.

Correct Answer: B) High-capacity networks fit noise before learning true features.

When expert demonstrations contain noisy or unintended movements, such as finger-twitching motions (biomechanical noise), adding more expert data to the training process can worsen the imitation learning (IL) behavior. This happens because high-capacity models, such as deep neural networks, have a tendency to overfit the noisy patterns present in the data before they learn the true underlying features of the task. The model starts to memorize the noise, interpreting it as part of the task, which leads to the policy exactly replicating unwanted noise in the expert's behavior. To mitigate this, regularization techniques or data augmentation strategies that reduce overfitting can be employed to help the model focus on the more relevant features of the task.

Q8

Is it possible to learn an imitation policy using only expert states (no action labels)? If yes, how?
A) No, expert actions are strictly required
B) Yes, using an inverse dynamics model
C) Yes, by guessing actions via optimization
D) Yes, but only in deterministic environments

Correct Answer: B) Yes, using an inverse dynamics model

It is indeed possible to learn an imitation policy using only expert states (without direct action labels) by leveraging an inverse dynamics model. In this approach, the inverse dynamics model is trained to predict the action taken by the expert given a pair of states (current state and next state). By learning the inverse relationship between states and actions, the agent can infer the expert's actions and then train a policy based on those inferred actions. This technique is particularly useful when action labels are unavailable or costly to obtain, as it enables the agent to mimic the expert's behavior solely from state transitions.