Principles of Robot Autonomy II
Exam 3
March 17, 2023

Name:

SUNet ID:

Instructions:

- Time allowed: 60 minutes.
- Total Points: 48.
- The exam consists of **three** problems.
- Please read all questions carefully before answering. Correct answers will receive full credit.
- **To get partial credit for an incorrect answer, you should explain your reasoning.** Please write all your answers in the provided answer sheet.

Good luck!

1. **Imitation Learning (24 points)**

(i) True or False: Unlike vanilla behavior cloning, the goal of DAgger is to find a policy that matches the feature expectations of the expert demonstrations.

(ii) You are trying to learn a linear reward function for an RC car that you want to move with a desired speed v_{des} when it is safe to do so. The linear reward function has the following form:

$$r(s) = \sum_i \omega_i \phi_i(s)$$

where s denotes the state, which includes the speed v of the vehicle along with other information, ω is the weight vector, and $\phi(s)$ is the features associated with state s. Assume only one of the features carries information about the speed v of the vehicle. One of your co-workers comes up with a list of possible functions for this speed feature to make the RC car move with v_{des}:

$$v, \qquad v - v_{\text{des}}, \qquad |v - v_{\text{des}}|, \qquad \mathbb{I}(v \geq v_{\text{des}}), \qquad \exp\left(-\frac{(v - v_{\text{des}})^2}{2}\right)$$

where $\mathbb{I}$ is the indicator function. Considering the fact that you are going to learn a linear reward function, which of the following features are reasonable for this task?

v

$v - v_{\text{des}}$

$|v - v_{\text{des}}|$

$\mathbb{I}(v \geq v_{\text{des}})$

$\exp\left(-\frac{(v-v_{\text{des}})^2}{2}\right)$

Explain:

(iii) True or False: Conditional imitation learning does not suffer from the problem of compounding errors.

(iv) True or False: In inverse reinforcement learning, there are many reward functions under which the expert demonstrations are optimal.

(v) True or False: In maximum margin planning, we can account for potential expert suboptimality by introducing a slack variable.

(vi) Which of the following is **not** an advantage of behavior cloning compared to reinforcement learning?

Removes the need for hand-designed reward functions

Improves the long-horizon planning

Avoids the problem known as reward hacking, in which reinforcement learning agents learn to exploit a reward function instead of the desired behavior

Explain:

Solution:

(i) False. Behavior cloning and DAgger aim to match the state-conditioned action distribution of the expert. Matching the feature expectations is the goal of some inverse reinforcement learning algorithms.

(ii) $|v - v_{\text{des}}|$ and $\exp\left(-\frac{(v-v_{\text{des}})^2}{2}\right)$ are reasonable, because $v = v_{\text{des}}$ either minimizes or maximizes them. For v and $v - v_{\text{des}}$, the car is going to learn either minimizing or maximizing the speed. For $\mathbb{I}(v \geq v_{\text{des}})$, the reward will be indifferent between the values that are larger (or smaller) than v_{des}.

(iii) False. Conditioning on the user goal alone will not address the problem of compounding error in imitation learning.

(iv) True. Reward ambiguity is an issue for IRL where multiple policies may lead to the same reward function.

(v) True.

(vi) Improves the long horizon planning. Behavior cloning does not perform long-horizon planning and actually suffers from compounding errors.

2. **Learning from Diverse Sources of Data (12 points)**

(i) True or False: Learned reward functions that are modeled with linear models are less expressive than those modeled with neural networks, because they consider only linear combinations of a predefined set of features.

(ii) True or False: In active preference-based learning, the agent chooses a pair of trajectories for a human to compare based on the human's responses to previous queries.

(iii) There are many forms of human feedback that robots can learn from, such as

- Offline expert demonstrations
- Interactive expert demonstrations
- Suboptimal demonstrations
- Physical corrections
- Pairwise comparisons of trajectories
- Large language models

Select one form of feedback from the list and answer the following questions:

- What is a pro and a con of this form of feedback?
- In one sentence, what is the key idea of the algorithm that leverages this form of feedback?
- Is the algorithm an example of direct policy learning or reward learning?

Explain:

Solutions:

(i) True.

(ii) True. The agent generates an informative query for the human to compare two trajectories given how the human has responded to the previous queries so far.

(iii) Various correct answers.

3. **Interaction-aware control and shared autonomy (12 points)**

(i) When modeling interactions between humans and robots in a game-theoretic fashion (theory of mind), we often struggle with the computational challenges of recursive belief modeling. Can you provide one way of addressing such computational challenges?

Explain:

(ii) An autonomous car with a limited-range camera is driving on a road that may have a 30-mph speed limit. It uses a POMDP to model the problem. Suppose the POMDP includes 1) two states $\mathcal{S} = \{0, 1\}$, representing whether the road has a 30-mph speed limit; 2) two observations $\mathcal{O} = \{0, 1\}$ representing whether the robot sees a speed limit sign; 3) a continuous action set $\mathcal{A} = [0, 60]$ representing the speed of the autonomous car. Assume the initial state distribution is uniform over $\mathcal{S}$. Let the transitions, observations and reward be modeled as:

$$P(s' \mid s, a) = \mathbb{I}(s' = s)$$

$$P(o \mid s) = \begin{cases} \frac{1}{2} & \text{if } s = 1, \\ 1 & \text{if } s = o = 0, \\ 0 & \text{otherwise.} \end{cases}$$

$$R(s, a, s') = -s(a - 30)^2 - (1 - s)(a - 60)^2$$

for $\forall s, s' \in \mathcal{S}, \forall o \in \mathcal{O}, \forall a \in \mathcal{A}$. What is the optimal action when $o = 0$?

0

20

30

50

60

Explain:

(iii) True or False: Q-MDP is a method that always gives the exact solution to POMDPs.

Solutions:

(i) Various correct answers, including approximating the game using a Stackelberg game (leader-follower game) and bringing down dimensionality of the action space by operating in the latent intent space.

(ii) (d). When $s = 0$, we have $o = 0$ with probability 1. When $s = 1$, we have $o = 0$ with probability 1/2. By Bayes' rule, $s = 0$ with probability 2/3 when $o = 0$. Since changing the state is not ever possible in this POMDP, we just need to maximize the immediate reward. Solving $a^* = \arg\max_a -\frac{2}{3}(a - 30)^2 - \frac{1}{3}(a - 60)^2$ gives $a^* = 50$.

(iii) False. QMDP only attempts to approximate the solution to POMDPs by assuming that full observability will be attained in the next time step.