Example Exam Questions

Reinforcement Learning

February 2022

Below you find 8 exemplary exam questions labelled from 'Easy' to 'Hard' to give you an idea of what the exam may look like. Each 'Easy' and 'Medium' question is worth 1, each 'Hard' question is worth 2 points. The exam will have a duration of 90 minutes and consist of ~ 25 questions and a total of 30 points, meaning that you will have around 3 minutes of time per point.

- (Easy) What are Upper-Confidence-Bounds in the context of K-armed bandit problems?

  Answer: Upper-Confidence-Bounds measure the 'potential' of a given bandit arm. Instead of just considering the arm with the highest expected value, the arm with the highest *potential* value is selected. This potential depends both the expected value and the estimated uncertainty of taking the action.

- (Easy) Briefly (2-3 Sentences) explain the moving target problem in DQNs.

  Answer: The target Q-Values in DQNs need to be fixed to prevent the targets of the loss function to move. If the targets are not fixed, we will change the targets by updating the parameters of the Q function. This leads to unstable training and an ill-defined loss function.

- (Easy) Name and briefly (1 sentence each) explain the 3 main paradigms/algorithm families in Imitation Learning.

  Answer:

  1. *Behavioral Cloning* learns a policy from expert data, similar to supervised learning.
  2. *Inverse Reinforcement Learning* recovers a reward function from expert data, and then learns a policy on this reward function
  3. *Adversarial Imitation Learning* trains a policy to produce data that is indistinguishable from the expert for a discriminator.

- (Medium) You are tasked to train an agent to play the ancient 2-player board game of Pai Sho. The game uses a discrete board state, with both players' actions consisting of placing or moving a limited number of tiles on the board.

  1. Which model-free and not planning-based algorithm from the lecture is best suited for this task? Explain your decision.
  2. Which planning-based method would you use instead? Explain your decision.

  Hint: As this is a combinatorial game, you may assume a fast simulator to be readily available. The number of states is very large.

  Answer:

  1. DQN is the algorithm of choice, since we are dealing with discrete action spaces, for which a Q-function is well-suited. We can not use tabular approaches because the state space is too large. *Note:* Any algorithm that trains an explicit policy would likely be less efficient because of the additional complexity of training both a Q/V Function and a policy instead of just the Q Function
  2. Monte Carlo Tree Search could be used to instead search the state space of the game. Combining this with a value and a policy function would result in an algorithm such as AlphaGo, which can plan multiple steps into the future and then estimate its value function and policy based on the resulting search tree.

- (Medium) Consider an optimization problem where the goal is to approximate a complex distribution $p^*(x)$ with a simple Gaussian $q(x)$.

  1. What kind of solution can you expect when using the forward KL-Divergence,

  $$\mathrm{KL}(p^*(x)||q(x)) = \int_x p^*(x) \log \frac{p^*(x)}{q(x)} \mathrm{d}x,$$

  and the reverse KL-Divergence,

  $$\mathrm{KL}(q(x)||p^*(x)) = \int_x q(x) \log \frac{q(x)}{p^*(x)} \mathrm{d}x.$$

  2. How can you see this kind of behavior from the equations of the forward and the reverse KL?

  Answer:

  1. For the forward KL-Divergence, the model will show mode averaging, i.e., it will try to cover as much probability mass of p^* as possible. For the reverse KL divergence, you can expect mode seeking behavior, i.e., the Gaussian to model one of the modes of p^*.
  2. In the forward KL, $q(x)$ must be large wherever $p^*(x)$ is large, causing it to average over all modes of $p^*(x)$. In the reverse KL, $q(x)$ must be close to zero wherever $p^*(x)$ is close to zero, but can be zero on modes of $p^*(x)$. This causes it to 'seek' modes.

- (Medium) in model-based Reinforcement Learning, the algorithm builds an explicit model of the world to help to maximize the cumulative reward.

  1. Briefly (1-2 sentences) explain why model-based RL is usually more sample efficient than approaches without a world model?
  2. What is the difference between a PoMDP and a regular MDP?
  3. Name 2 common failure cases for model-based RL.

  Answer:

  1. Having a world model allows for exploration and planning in a learned space, i.e., without interaction with the actual environment. As such, the policy can improve through the world model rather than the real world, which heavily reduces the sample complexity.
  2. In a PoMDP, the state $\mathcal{S}$ can only be observed through an observation space $\mathcal{O}$. The observation model $p(\mathbf{o}_t|\mathbf{s}_t)$ can lose information, i.e., the observation can be incomplete.
  3. (a) The model may be over-optimistic in certain regions of the state space, leading the policy to exploit problems with the model rather than learning a policy that is useful in the real world.
     (b) If there is a lack of exploration in the policy, the model may not cover all the relevant regions of the state space. The model error can then be very low, even though the model is not useful.

- (Hard) You are given the following optimization problem:

$$\begin{aligned} &\mathrm{argmax}_{\pi(a)} \int_a \pi(a) R(a) \mathbf{d}a + \alpha \mathrm{H}(\pi(a)) \\ &\text{s.t. } \mathrm{H}(\pi(a)) \geq \epsilon \\ &\int_a \pi(a) \mathrm{d}a = 1, \end{aligned}$$

  where α is a scalar and $\mathrm{H}(\pi(a)) = -\int_a \pi(a) \log \pi(a) \mathrm{d}a$ is the entropy.

  1. Write down the Lagrangian.
  2. Derive the optimal $\pi^*(a)$ depending on your Lagrangian multipliers.

3. Which condition does each of the Lagrangian multipliers need to satisfy for optimizing the dual function?

Hint 1: Note that we have a maximization problem here! Therefore, take care of the signs for your constraint when writing down the Lagrangian.

Hint 2: You can assume that fractions are well-defined.

Answer:

1. The Lagrangian is

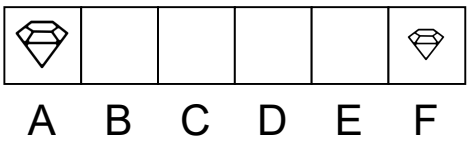
$$L(\pi(a), \eta, \lambda) = \int_a \pi(a)R(a)\mathbf{d}a - \alpha \int_a \pi(a)\log\pi(a)\mathbf{d}a - \eta(\epsilon + \int_a \pi(a)\log\pi(a)\mathbf{d}a) + \lambda(\int_a \pi(a)\mathbf{d}a - 1)$$

2. Deriving the optimal $\pi^*(a)$ gives

$$\begin{aligned} \frac{dL}{d\pi} &= R(a) - (\alpha + \eta)(\log\pi(a) + 1) + \lambda \\ \log\pi(a)(\alpha + \eta) &= R(a) - \alpha - \eta + \lambda \\ \pi^*(a) &= \exp(\frac{R(a) - \alpha - \eta + \lambda}{\alpha + \eta}) \end{aligned}$$

3. The Lagrangian Multiplier η needs to be ≥ 0. The Lagrangian Multiplier λ can be a positive as well as a negative number.

- (Hard) In the following MDP, your actions are going left ($\leftarrow$), going right ($\rightarrow$) and exit ($\times$). The only non-zero rewards are rewards for finding a goal state, which are $r(s = A, a = \times) = 6$ and $r(s = F, a = \times) = 1$, as well as a small penalty for moving to the left in state C, which is given by $r(s = C, a = \leftarrow) = -1$.



For which value of $\gamma > 0$ are choosing $\leftarrow$ and $\rightarrow$ equally good in state D?

Hint 1: First think about how the optimal action sequences from state D look like to the left and to the right. Afterwards, calculate the return $G = \sum_t \gamma^t r_t$ for both ways.

Hint 2: A solution to an equation of the form $0 = ax^2 + bx + c$ is given as

$$x_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Answer:
The optimal action sequence from D towards state F is $\rightarrow, \rightarrow, \times$ $G_1 = \gamma^0 r(s = D, a = \rightarrow) + \gamma^1 r(s = E, a = \rightarrow) + \gamma^2 r(s = F, a = \times) = \gamma^2$ (0.25P)

The optimal action sequence from state D towards state A is $\leftarrow, \leftarrow, \leftarrow, \times$ (0.5P)

$G_2 = \gamma^0 r(s = D, a = \leftarrow) + \gamma^1 r(s = C, a = \leftarrow) + \gamma^2 r(s = B, a = \leftarrow) + \gamma^3 r(s = A, a = \times) = -\gamma + 6 \cdot \gamma^3$

Set $G_1 = G_2$ and solve for γ

$$\gamma^2 = -\gamma + 6 \cdot \gamma^3 \tag{1}$$
$$0 = -\gamma - \gamma^2 + 6\gamma^3 \tag{2}$$
$$0 = -1 - \gamma + 6\gamma^2 \tag{3}$$
$$x \in \{-\frac{1}{3}, \frac{1}{2}\} \tag{4}$$
$$\rightarrow x = \frac{1}{2} \tag{5}$$