Technische Universität Darmstadt
FG Kommunikationstechnik
Prof. Dr.-Ing. Anja Klein
Dr. rer. nat. Sabrina Klos and Dr.-Ing. Andrea Ortiz

Fundamentals of Reinforcement Learning

Theory Exercise 1: Solution

Task 1 – Moodle Discussion: Reinforcement Learning Tasks

Potential RL tasks are those where a an active decision-making agent interacts with its environment and tries to achieve a goal despite uncertainty about its environment.
Note that when modeling your own RL task, it is crucial to design effective reward signals. In RL, the agent learns to maximize its reward. So if we want an agent to achieve a goal for us, we must provide rewards to the agent in such a way that in maximizing the rewards, the agent will actually achieve our goal. Hence, we need to make sure that the rewards we design truly indicate what we want to be accomplished. In particular, while the reward signal should communicate to the agent *what* we want to be achieved, it should not be used to give the agent prior knowledge about *how* to achieve what we want it to do (e.g., a chess-playing agent should be rewarded only for actually winning, not for achieving subgoals such as taking its opponent's pieces or gaining control of the center of the board).
Moreover, rewards are scalar feedback signals. In some cases, however, we might have a conceptual idea of what should be rewarded (e.g., an agent recommending me to switch off my phone should be rewarded if this increases my productiveness in studying). In such a case, we have to translate this conceptual idea of what should be rewarded to some numerical reward signal (e.g., find a measure for productiveness in studying and give agent reward proportional to this measure).
Note also that the agent's actions are the decisions the agent should learn how to make. So, specifically, it is under the control of the agent to select those actions. To properly model an RL task, what we model as actions has to be some decision that the agent can take. Anything that cannot be changed arbitrarily by the agent (e.g., an agent may not be able to *enforce* a person to turn off the phone) cannot be considered an agent's action, it is instead considered to be outside and thus part of the agent's environment. We should in such a case clearly define what the agent is able to decide on its own (e.g., agent may decide to *recommend* a person to turn off the phone).

Task 2 - Moodle Discussion: The Reward Hypothesis

There is no right or wrong answer in giving your opinion on the reward hypothesis. The goal of this task is to discuss the reward hypothesis since it is one of the most important characteristics of reinforcement learning to use a reward signal to formalize the idea of a goal. Although formulating goals in terms of reward signals might appear to be limiting, in practice it is widely applicable. However, note as discussed under Task 1, that if we want an agent to achieve a goal for us, we must provide rewards to it in such a way that in maximizing them the agent will also achieve our goals. More opinions and discussion on the reward hypothesis can be found in a homepage by Richard Sutton[1].

Task 3: Basic Probability

First, we define the three events

$$\begin{aligned} \mathrm{D} &= \{\text{Laptop is defective}\} \\ \mathrm{A} &= \{\text{Laptop was produced at location A}\} \\ \mathrm{B} &= \{\text{Laptop was produced at location B}\}. \end{aligned}$$

We are given that

$$\mathbb{P}(\mathrm{D}|\mathrm{A}) = 0.15 \qquad \text{and} \qquad \mathbb{P}(\mathrm{D}|\mathrm{B}) = 0.05.$$

Moreover, we know from the problem assignment that 1 000 000 laptops were produced at location A and 150 000 at location B. Thus, the total number of laptops is $1\,000\,000 + 150\,000 = 1\,150\,000$ and we obtain

$$\mathbb{P}(\mathrm{A}) = \frac{1\,000\,000}{1\,150\,000} \qquad \text{and} \qquad \mathbb{P}(\mathrm{B}) = \frac{150\,000}{1\,150\,000}.$$

Finally, we can compute the probability $\mathbb{P}(D)$ of purchasing a defective laptop by applying the law of total probability:

$$\begin{aligned} \mathbb{P}(D) &= \mathbb{P}(\mathrm{D}|\mathrm{A})\mathbb{P}(\mathrm{A}) + \mathbb{P}(\mathrm{D}|\mathrm{B})\mathbb{P}(\mathrm{B}) \\ &= 0.15 \cdot \frac{1\,000\,000}{1\,150\,000} + 0.05 \cdot \frac{150\,000}{1\,150\,000} \\ &= 0.137. \end{aligned}$$

The probability of purchasing a defective laptop is 13.7%.

[1] http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html

Task 4: Distributions of Discrete Random Variables

4.1)

The sample space is given by

$$\begin{aligned}\Omega &= \{\text{N,S,E,W}\} \times \{\text{N,S,E,W}\} \times \{\text{N,S,E,W}\} \times \{\text{N,S,E,W}\} \\ &= \{\text{N,S,E,W}\}^4,\end{aligned}$$

where N,S,E and W correspond to the directions north, south, east and west, respectively.

4.2)

The random variable X is defined as

$$X(\omega) = \text{Number of steps along the vertical axis in the sequence } \omega.$$

The number of steps along the vertical axis in one sequence ω can lie between 0 and 4, since one sequence consists of four steps. Thus, X can take the values $0, 1, 2, 3, 4$, i.e.,

$$X : \Omega \to \{0, 1, 2, 3, 4\}.$$

4.3)

We compute the PMF $f_X(x) = \mathbb{P}(X = x)$ by determining the probabilities for X to take one of the values $\{0, 1, 2, 3, 4\}$. For each step in the sequence, the probabilities of making a step along the vertical axis and making a step along the horizontal axis are both 0.5. Thus, the probability of making all four steps along either the horizontal or vertical axis is

$$f_X(0) = \mathbb{P}(X = 0) = \overbrace{\underbrace{\frac{1}{2}}_{\mathbb{P}(\text{horizontal step})} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}}^{\text{four independent steps}} = \frac{1}{16}$$

$$f_X(4) = \mathbb{P}(X = 4) = \overbrace{\underbrace{\frac{1}{2}}_{\mathbb{P}(\text{vertical step})} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}}^{\text{four independent steps}} = \frac{1}{16}.$$

For the remaining three probabilities, one has to take into account that the steps along the vertical axis can occur at different positions within the sequence. Thus, for one, two or three vertical steps, there are multiple possible sequences. By using

the binomial coefficient we obtain

$$f_X(1) = \mathbb{P}(X=1) = \underbrace{\binom{4}{1}}_{\text{1 out of 4 steps is vertical}} \cdot \underbrace{\frac{1}{2}}_{\mathbb{P}(\text{1 vertical step})} \cdot \underbrace{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}}_{\mathbb{P}(\text{3 horizontal steps})} = \frac{1}{4}$$

$$f_X(2) = \mathbb{P}(X=2) = \underbrace{\binom{4}{2}}_{\text{2 out of 4 steps are vertical}} \cdot \underbrace{\frac{1}{2}\cdot\frac{1}{2}}_{\mathbb{P}(\text{2 vertical steps})} \cdot \underbrace{\frac{1}{2}\cdot\frac{1}{2}}_{\mathbb{P}(\text{2 horizontal steps})} = \frac{3}{8}$$

$$f_X(3) = \mathbb{P}(X=3) = \underbrace{\binom{4}{3}}_{\text{3 out of 4 steps are vertical}} \cdot \underbrace{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}}_{\mathbb{P}(\text{3 vertical steps})} \cdot \underbrace{\frac{1}{2}}_{\mathbb{P}(\text{1 horizontal step})} = \frac{1}{4}.$$

We can also find the PMF based on the following observation:

The random variable X in our example can be understood as the number of heads in a sequence of four fair and independent coin tosses, as taking a vertical step and throwing a head both have the probability 0.5. The number of heads in a sequence of four fair and independent coin tosses follows a Binomial distribution $\text{Bin}(n,p)$ with $n=4$ and $p=0.5$. Hence,

$$X \sim \text{Bin}(4, 0.5),$$

which gives us exactly the same PMF as computed above.

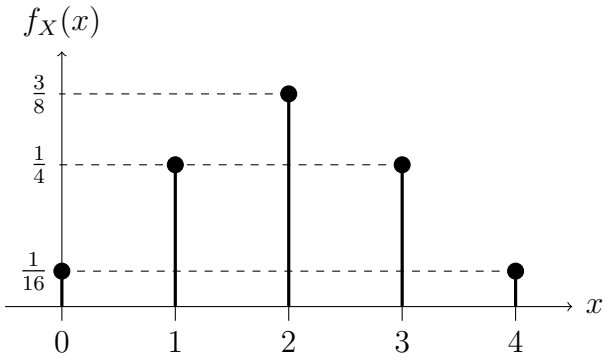

Figure 1: PMF of random variable X

For computing the CDF $F_X(x) = \mathbb{P}(X \leq x)$, we add up the probabilities appearing

in the PMF:

$$F_X(0) = \mathbb{P}(X \leq 0) = f_X(0) = \frac{1}{16}$$
$$F_X(1) = \mathbb{P}(X \leq 1) = f_X(0) + f_X(1) = \frac{5}{16}$$
$$F_X(2) = \mathbb{P}(X \leq 2) = f_X(0) + f_X(1) + f_X(2) = \frac{11}{16}$$
$$F_X(3) = \mathbb{P}(X \leq 3) = \sum_{i=0}^{3} f_X(i) = \frac{15}{16}$$
$$F_X(4) = \mathbb{P}(X \leq 4) = \sum_{i=0}^{4} f_X(i) = 1.$$

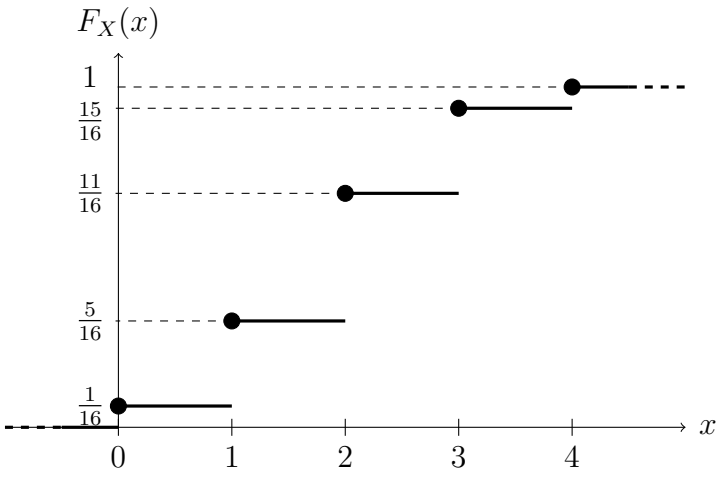

Figure 2: CDF of random variable X

Task 5: The Zipf Distribution

5.1)

When the video request process follows the Zipf distribution with parameter α, the relative probability of a request for the i-th most popular video is proportional to $\frac{1}{i^\alpha}$, so inversely proportional to its rank to the power of α.

5.2)

For $\alpha = 0$, the Zipf distribution corresponds to a uniform distribution, i.e., the request probability is equal for all videos. As α grows, the more skewed the Zipf distribution becomes, i.e., the parameter α determines the skewness of the distribution.

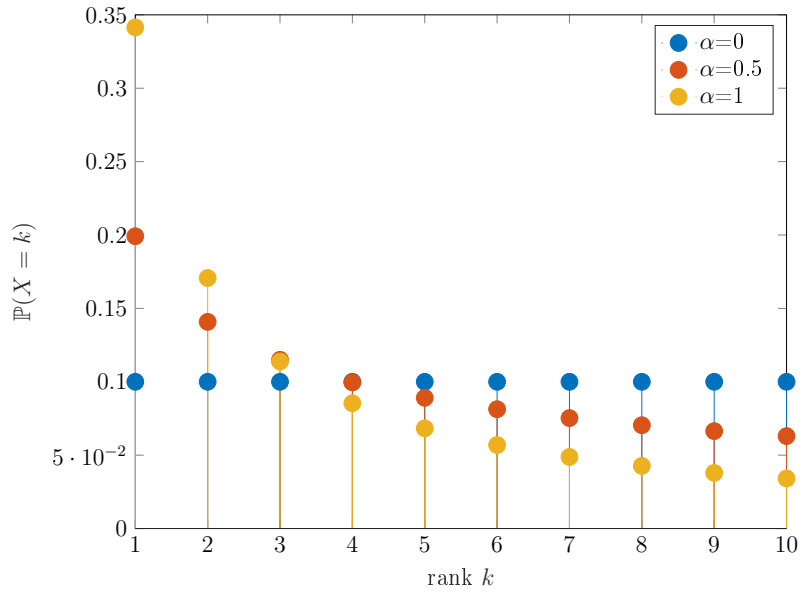

Figure 3: PMF of Zipf distribution for different α and $N = 10$

5.3)

The cache size m should be selected in such a way that

$$\mathbb{P}(X \leq m) \geq 0.25,$$

where $X \leq m$ means that one of the m most popular videos is requested. Note that those videos are the ones stored in the cache. By exploiting the hint given in the problem assignment, for $N = 100$ and $\alpha = 1$ we have

$$\mathbb{P}(X = 1) = \frac{\frac{1}{1}}{\sum_{n=1}^{100} \frac{1}{n}} \approx \frac{1}{5.1874} = 0.1928$$

$$\mathbb{P}(X = 2) = \frac{\frac{1}{2}}{\sum_{n=1}^{100} \frac{1}{n}} \approx \frac{1}{2 \cdot 5.1874} = 0.0964.$$

We observe that $\mathbb{P}(X \leq 2) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = 0.2892 \geq 0.25$. Thus, the content provider should select a cache size of at least $m = 2$ videos.