Homework 3

Due October 8 at 11 pm

1. (Fish) A biologist is studying a rare species of fish. She captures four individuals and measures their weights, which are 5, 8, 5 and 6 kg.

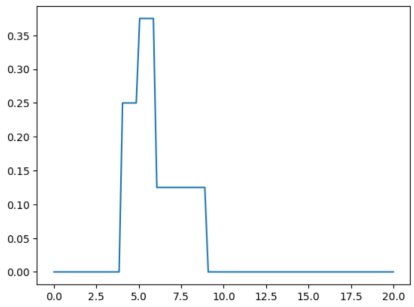
   (a) What is the empirical conditional probability that a fish weighs more than 7 kg given that they weigh more than 6 kg?

   The empirical conditional probability we want to find is:
   $\frac{\textbf{Number of fish with weights bigger than 7kg}}{\textbf{Number of fish with weights bigger than 6kg}} = \frac{1}{1} = 1$

   (b) Plot an estimate of the pdf of the fish weight using kernel density estimation with a rectangular kernel of width 2.

   The rectangular kernel is defined as $K(a, x_i, h) = \begin{cases} \frac{1}{2} & a \in [x_i - h/2, x_i + h/2] \\ 0 & otherwise \end{cases}$

   The plot and codes are shown below:




```python
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt
import seaborn as sns

def rect_kernel(data, center, width):
    return 1/width if data <= center
    + width/2 and data >= center - width/2 else 0

def compute_kde(data_points, t):
    return  1 / len(data_points)
    * sum([rect_kernel(t, data_point, 2)
    for data_point in data_points])

indices = np.linspace(0,20,100)
plt.plot(indices, [compute_kde(data_points, index)
for index in indices])
```

(c) What is the conditional probability that a fish weighs more than 7 kg given that they weigh more than 6 kg according to your estimated pdf?

The probability is calculated as $\frac{P(X \geq 7)}{P(X \geq 6)} = \frac{0.5}{0.875} \approx 0.5714285714285714$. The code is attached below:

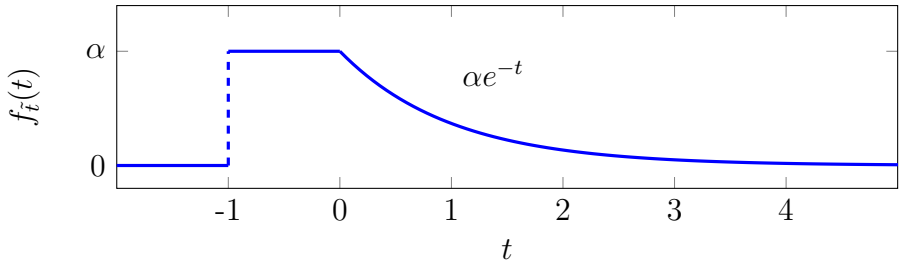
```
def rect_kernel(data, center, width):
    return 1/width if data <= center + width/2 and data
    >= center - width/2 else 0

def compute_kde(data_points, t):
    return  1 / len(data_points)
    * sum([rect_kernel(t, data_point, 2)
    for data_point in data_points])

def compute_prob(data_points, low, high):
    return sum([compute_kde(data_points, t)
    for t in range(low, high + 1)])

compute_prob(data_points, 7, 20) /
compute_prob(data_points, 6, 20) -> 0.5714285714285714
```

2. (Nuclear power plant) The random variable $\tilde{t}$ with the following pdf



models the time at which there is a leak in a nuclear power plant. The pdf is constant during the time the station is built (between -1 and 0) and exponential with parameter 1 afterwards (from 0 to $+\infty$).

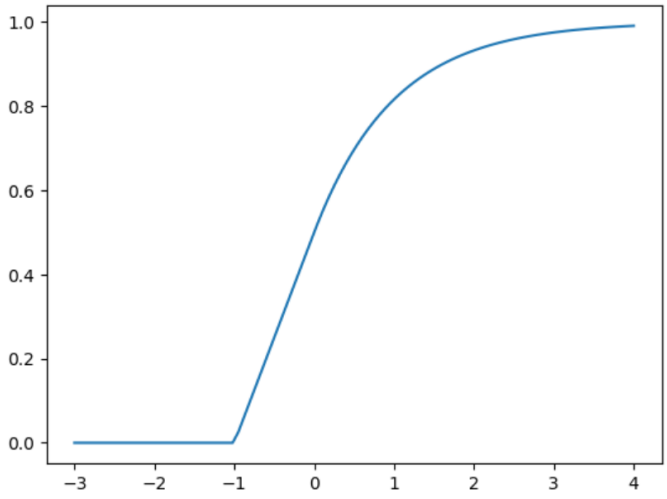
(a) Compute the value of the constant α.

The pdf is given as $f_{\tilde{t}}(t) = \begin{cases} \alpha & t \in [-1, 0] \\ \alpha e^{-t} & t > 0 \end{cases}$. In order for this to be a valid pdf, we must check the following: $\int_{-1}^{0} \alpha dt + \int_{0}^{\infty} \alpha e^{-t} dt = 2\alpha = 1$ where $\alpha = \frac{1}{2}$.

(b) Compute the cdf of $\tilde{t}$ and plot it.

$$F_{\tilde{t}}(t) = \int_{-1}^{\infty} f_{\tilde{t}}(t)dt = \begin{cases} 0 & t < -1 \\ \int_{-1}^{x} \frac{1}{2}dt & t \in [-1,0] \\ \int_{-1}^{0} \frac{1}{2}dt + \int_{0}^{t} \frac{1}{2}e^{-t}dt & t > 0 \end{cases} = \begin{cases} 0 & t < -1 \\ \frac{1}{2}t + \frac{1}{2} & t \in [-1,0] \\ 1 - \frac{1}{2}e^{-t} & t > 0 \end{cases}$$

The plot and codes are shown as follows:




```
def piecewise_function(t):
    if t < -1:
        return 0
    elif -1 <= t <= 0:
        return 0.5 * t + 0.5
    else:
        return 1-0.5*np.exp(-t)

x_axis = np.linspace(-3,4,100)

plt.plot(x_axis,[piecewise_function(t) for t in x_axis])
```

(c) Compute the pdf of $\tilde{t}$ conditioned on $\tilde{t} < 0$.

We first compute the conditional CDF:

$$F_{\tilde{t}|\tilde{t}<0}(t) = \frac{P(\tilde{t}\leq t \cap \tilde{t}<0)}{P(\tilde{t}<0)} = \begin{cases} \frac{P(\tilde{t}<0)}{P(\tilde{t}<0)} & t > 0 \\ \frac{P(\tilde{t}\leq t)}{P(\tilde{t}<0)} & -1 \leq t \leq 0 \\ 0 & t < -1 \end{cases} = \begin{cases} 1 & t > 0 \\ \frac{\int_{-1}^{t} \frac{1}{2}dt}{\int_{-1}^{0} \frac{1}{2}dt} = t+1 & -1 \leq t \leq 0, \text{ then} \\ 0 & t < -1 \end{cases}$$

we take the gradient of it and could get that $f_{\tilde{t}|\tilde{t}<0}(t) = \begin{cases} 1 & -1 \leq t \leq 0 \\ 0 & otherwise \end{cases}$

3. (Measurements) You have access to the readings of a device that indicates whether a radioactive particle has decayed. However you do not get a continuous reading, you get a reading every second.

   (a) A reasonable model for the time the particle takes to decay is that it is a random variable with pdf

$$f_{\tilde{t}}(t) := \begin{cases} \lambda \exp(-\lambda t), & \text{if } t \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

   where λ is a fixed constant. Taking into account that the measurement device rounds up the time and outputs an integer number of seconds,compute the pmf of the reading from the device. What kind of random variable is this?

   Denote the reading of the device by X and the time that passed for particle decaying by T, we have the following relationship between X and T: If $X = x$ then $T \in (x-1, x]$. Thus we can compute the PMF of X first as follows:

$$\begin{aligned} P_X(x) &= P(X = x) \\ &= P(T \in (x-1, x]) \\ &= \int_{x-1}^{x} f_T(t)dt \\ &= e^{-\lambda(x-1)} - e^{-\lambda x} \\ &= (e^{-\lambda})^{x-1}(1 - e^{-\lambda}) \end{aligned}$$

   ,$\forall X = 1, 2, 3, \cdots$.Also, the $P(X = 0) = P(T = 0) = 0$ since T is a continuous random variable. Thus by the PMF we got , we have found that $X \sim Geo(1 - e^{-\lambda})$.

   (b) What is the pdf of the error between your reading and the true time of decay?

   We denote the CDF of the error as E and the CDF of it is:

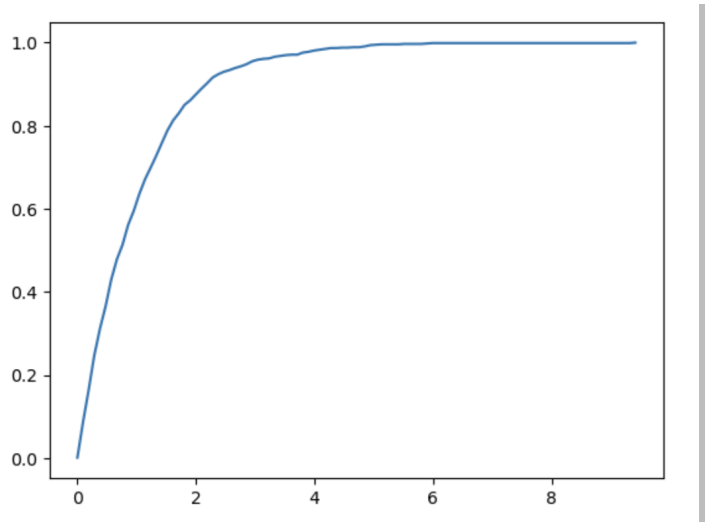
$$\begin{aligned} F_E(\epsilon) &= P(E \leq \epsilon) \\ &= \sum_{x=1}^{\infty} P(x - \epsilon \leq T \leq x) \\ &= \sum_{x=1}^{\infty} \left(\int_{x-\epsilon}^{x} \lambda e^{-\lambda t} dt \right) \\ &= \sum_{x=1}^{\infty} e^{-\lambda(x-\epsilon)} - e^{-\lambda x} \\ &= (e^{\lambda\epsilon} - 1) \sum_{x=1}^{\infty} e^{-\lambda x} \\ &= (e^{\lambda\epsilon} - 1) \frac{e^{-\lambda}}{1 - e^{-\lambda}} \qquad \text{(Geometric series)} \end{aligned}$$

   , we then take the derivative with respective to ϵ and get : $f_E(\epsilon) = \frac{\lambda e^{\lambda\epsilon}}{e^{\lambda} - 1}, 0 < \epsilon < 1$ and 0 otherwise.

4. (Applying the cdf) The array in `samples.npy` contains $n := 1,000$ i.i.d. samples from a certain distribution.

   (a) Compute the empirical cdf of the data F_X and plot it.

   The empirical CDF for the data is just $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \le x\}$. The plot and codes are as follows:




```
def ecdf(data_points, t):
    return sum(np.array(data_points) <= t)
    / len(data_points)

x_axis = np.linspace(np.min(data),np.max(data),100)
plt.plot(x_axis, [ecdf(data, t) for t in x_axis])
```

   (b) If you apply the empirical cdf to each data point x_i , $1 \le i \le n$, to obtain a new data point $y_i := F_X(x_i)$, what are the new data equal to? Does your answer depend on the distribution of the data?

   The new data should equal to $\{\frac{1}{n}, \frac{2}{n}, \cdots, \frac{n-1}{n}, 1\}$, which doesn't depend on the distribution of dataset.