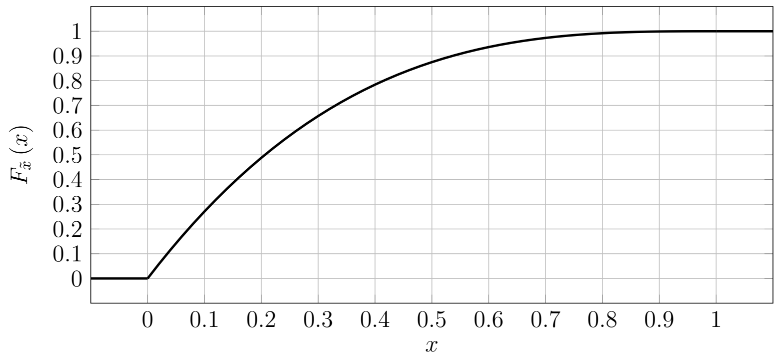
Midterm practice problems

1. (Cumulative distribution function) The random variable $\tilde{x}$ has the following cdf.



   (a) What is the median of $\tilde{x}$?

   (b) Is the probability density at 0.2 higher or lower than at 0.8? (No need to calculate anything, just look at the cdf)

   (c) If we know that $\tilde{x} \geq 0.2$, what is the conditional probability that $\tilde{x}$ is smaller than 0.4?

   (d) You have access to the following three samples from a uniform distribution: 0.2, 0.4, 0.55. Use them to simulate three samples from $\tilde{x}$.

2. (Fire alarm) We are interested in analyzing a newly-installed fire alarm, which can behave strangely if the battery malfunctions. We model the alarm with a Bernoulli random variable $\tilde{a}$ ($\tilde{a} = 1$ if the alarm rings, $\tilde{a} = 0$ if it doesn't), fire with another Bernoulli random variable $\tilde{x}$ ($\tilde{x} = 1$ if there is fire, $\tilde{x} = 0$ if there is no fire), and the behavior of the battery with a third Bernoulli random variable $\tilde{b}$ ($\tilde{b} = 1$ if the battery malfunctions, $\tilde{b} = 0$ if it is fine). After gathering some data, we determine that the probability of the alarm ringing depends on the fire alarm and the state of the battery in the following way:

	No fire	Fire
Battery is fine	0	0.8
Battery malfunctions	0.2	0.4

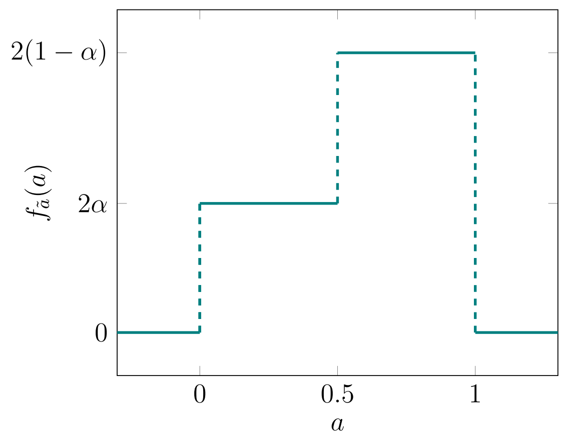
   To be clear, the table contains the probability of $\tilde{a} = 1$ conditioned on all possible values of $\tilde{x}$ and $\tilde{b}$. In all your calculations assume that the occurrence of fire and the state of the battery are independent (i.e. $\tilde{x}$ and $\tilde{b}$ are independent). The probability of a fire is 0.2 and the probability of the battery malfunctioning is 0.5.

   (a) Compute the marginal pmf of $\tilde{a}$.

   (b) The alarm rings. Compute the conditional joint pmf of $\tilde{x}$ and $\tilde{b}$ given this event.

(c) Are $\tilde{x}$ and $\tilde{b}$ conditionally independent given $\tilde{a}$? Justify your answer mathematically and intuitively.

3. (Rat) We are interested in modeling the location of a rat that lives in a tunnel in the New York subway. The tunnel is 1 mile long. The first half mile is very dark. The second half is not as dark but has more food. We decide to use a parametric model where the pdf of the location of the rat is equal to 2α over $[0, 0.5]$, and to $2(1-\alpha)$ over $[0.5, 1]$, where α is a parameter. Here is a plot of the pdf:



(a) What two conditions must a function satisfy to be a valid pdf? For what values of α does this function satisfy them?

(b) Compute the maximum likelihood estimate of α if we observe the following independent samples from the proposed model: 0.1, 0.8, 0.9, 0.7, 0.3.

(c) What is the probability that the rat is in the first half of the tunnel? Based on your answer, suggest a way to estimate α based on the empirical probability estimator. Compare it to the maximum-likelihood estimator for α.

(d) Use kernel density estimation with a rectangular kernel of width 0.2 to approximate the probability density using the same data as in (b). What is problematic about the estimate you obtain. How would you alleviate this?

4. (Election) The made-up country Lalaland is composed of three states. Their presidential election is won by the candidate that has a majority in more states. We model the result of a Lalaland presidential election with two candidates (A and B) as a Bernoulli random variable $\tilde{o}$ ($\tilde{o} = 1$ means that A wins). We model the result in each state as Bernoulli random variables $\tilde{s}_1$, $\tilde{s}_2$ and $\tilde{s}_3$ ($\tilde{s}_i = 1$ means that A wins state i). From the available data we determine that the probability that candidate A wins state 1, 2, or 3 is equal to the same value: 0.6. In addition, we assume that $\tilde{s}_1$, $\tilde{s}_2$ and $\tilde{s}_3$ are mutually independent.

(a) What is the probability that candidate A wins the election?

(b) If B wins the election, what is the conditional probability that they won in state 2?

(c) Are $\tilde{s}_1$ and $\tilde{s}_2$ conditionally independent given $\tilde{o}$? Justify your answer mathematically and also explain it intuitively.

5. (Earthquake) A geophysicist is trying to estimate the pdf of the inter-arrival times of earthquakes in a certain region. The available data (in years) are: 7.5, 10, 32.5.

   (a) If she estimates the density using kernel density estimation with a rectangular kernel of width 10, what is the probability that the inter-arrival time is larger than 10?

   (b) If she estimates the density by applying maximum likelihood estimation with an exponential parametric model, what is the probability that the inter-arrival time is larger than 10?

   (c) Briefly describe the advantages and disadvantages of the nonparametric and parametric models.

6. (Scaling random variables) Are these statements true or false? Justify your answers mathematically.

   (a) If $\tilde{a}$ is exponential with parameter λ, for any $\alpha > 0$, the random variable $\tilde{b} := \alpha\tilde{a}$ is exponential with parameter $\alpha\lambda$.

   (b) If $\tilde{a}$ is geometric with parameter θ, for any positive integer α, the random variable $\tilde{b} := \alpha\tilde{a}$ is geometric with parameter $\alpha\theta$.

   (c) If the median of a random variable $\tilde{a}$ is m, then the median of $\tilde{b} := \alpha\tilde{a} + \beta$ is $\alpha m + \beta$ for any $\alpha > 0$ and $\beta \in \mathbb{R}$.

7. (Baby) Anna is having a baby, which will be a boy or a girl with probability 1/2. When the baby is born, she will call her aunt Margaret to tell her whether the baby is a boy or a girl. Margaret is a bit deaf; she will misunderstand and think the baby is the wrong sex with probability 0.2. Then Margaret will tell her neighbor Bob, who is also a bit deaf. He will misunderstand what Margaret says (i.e. he will think that he baby is the opposite sex of what she says) with probability 0.1.

   (a) What is the probability that both Margaret and Bob think that the baby is a girl?

   (b) If Bob thinks the baby is a girl, what is the probability that he is right?

   (c) Is what Bob thinks conditionally independent from the sex of the baby, given what Margaret thinks?

   (d) Is what Bob thinks independent from the sex of the baby?

8. (Missing data) A medical researcher is trying to determine the probability that a certain drug produces side effects. Unfortunately, for some patients it is unknown whether there were any side effects, the data are missing because they never returned to the doctor. Here is the dataset:

	Side effects	No side effects	Missing
Men	40	10	50
Women	30	60	10

   (a) A common assumption when dealing with missing data is that the data are *missing at random*, meaning that each data point is missing with the same probability, independently from all the others. Do you think this holds here?

(b) What is the probability of a patient suffering side effects according to the observed data?

(c) Assuming that side effects are conditionally independent from the data being observed given the patient sex, estimate the true probability that a patient suffers side effects.