Midterm practice problems

1. (Short questions)

   (a) Let $\tilde{x}$, $\tilde{y}$ and $\tilde{z}$ be arbitrary random variables. Does this always hold? (Only justify your answer if it is *Yes.*)

   $$\int_{y=-\infty}^{\infty} f_{\tilde{x}\,|\,\tilde{y},\tilde{z}}\,(x\,|\,y,z)\;\mathrm{d}y = 1. \tag{1}$$

   (b) Let $\tilde{x}$, $\tilde{y}$ and $\tilde{z}$ be arbitrary random variables. Does this always hold? (Only justify your answer if it is *Yes.*)

   $$\int_{x=-\infty}^{\infty} f_{\tilde{x}\,|\,\tilde{y},\tilde{z}}\,(x\,|\,y,z)\;\mathrm{d}x = 1. \tag{2}$$

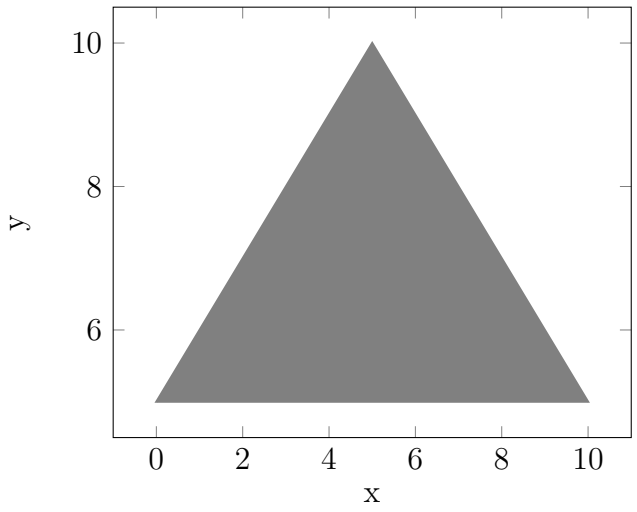
   (c) Is this identity true? Justify your answer without computing anything.

   $$\sum_{k=0}^{20} \int_{u=0}^{1} \binom{20}{k} u^k\,(1-u)^{(20-k)}\;\mathrm{d}u = 1. \tag{3}$$

   (d) The random variables $\tilde{x}$, $\tilde{y}$ and $\tilde{z}$ take the values in the set $\{1, 2, 3\}$. How many parameters do we need to specify their joint pmf? How many parameters do we need if $\tilde{x}$ and $\tilde{y}$ are independent so that we can use the factorization $p_{\tilde{x}} p_{\tilde{y}} p_{\tilde{z}\,|\,\tilde{x},\tilde{y}}$?

   (e) Let $\tilde{u}$ be uniformly distributed between 0 and 1. What is the distribution of $\tilde{u}$ conditioned on the event $\tilde{u} > 1/2$?

   (f) Two random variables $\tilde{x}$ and $\tilde{y}$ have a joint pdf $f_{\tilde{x},\tilde{y}}(x, y)$ that is nonzero at every point in the shaded region:

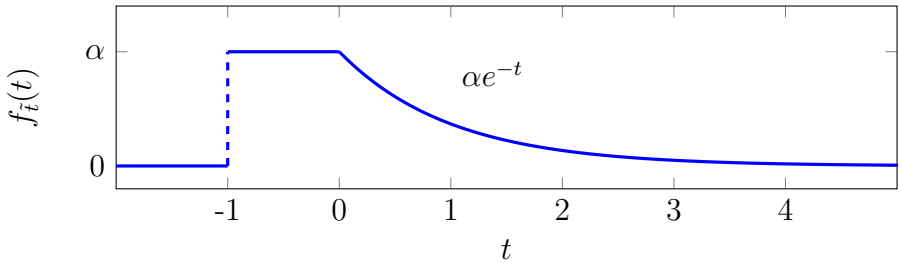


   In what region is the joint cdf $F_{\tilde{x},\tilde{y}}(x, y)$ equal to 0? In what region is it equal to 1?

(g) A biologist is interested in modeling the probability that each tree in a forest produces flowers. She models each of the 100 trees in the forest as a random variable $\tilde{x}_i$, $1 \leq i \leq 100$, that equals 1 if the tree produces flowers and 0 if it does not. She reasons that the random variables are very dependent on each other: the trees share the same type of soil, they receive similar amounts of sunshine and rain, etc. She decides to model the joint distribution of the 100 random variables, and estimate it from past data. Is this possible without making any independence or conditional independence assumptions?

(h) The price of flights to Greece, and the price of hotels in Greece increase a lot in the summer. If we estimate the yearly variance of the cost of a trip to Greece by adding the variance of the flight price and the variance of the hotel price, do you think we will get a good estimate of the total variance? Explain.

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 constant 1 afterwards (from 0 to $+\infty$).

(a) Compute the value of the constant α.

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

(c) Given a sample of a uniform random variable equal to 0.3, simulate a sample of $\tilde{t}$.

(d) Express $\tilde{t}$ as the mixture of two common parametric models. Define the pmf of the discrete random variable governing the mixture, as well as the two conditional parametric pdfs.

(e) Compute the mean of $\tilde{t}$.

(f) Compute the variance of $\tilde{t}$.

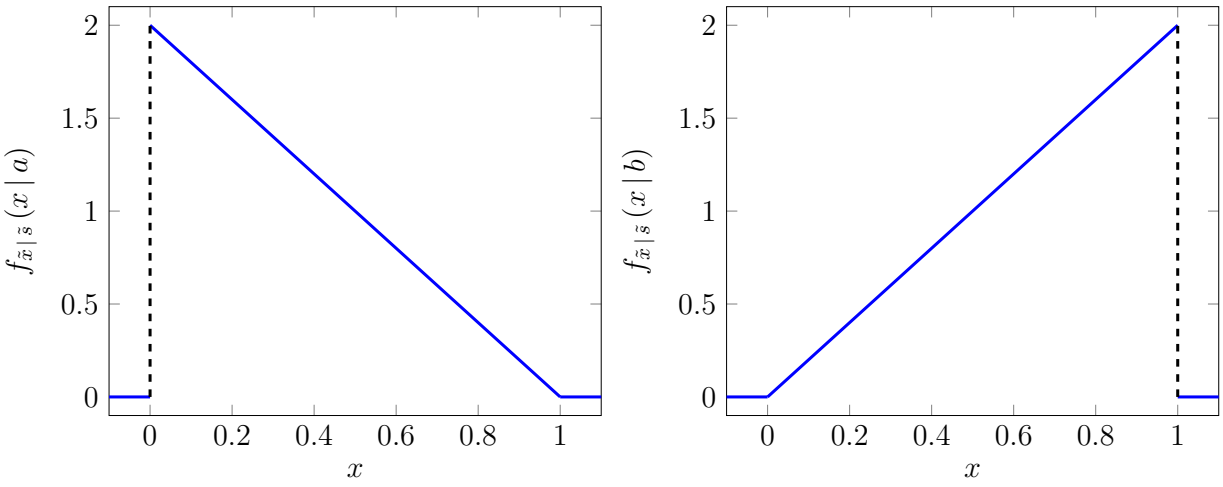
3. (Noisy data) A signal $\tilde{x}$ is equal to -1 or 1 with probability 1/2. When it is measured, the measurement $\tilde{y}$ equals $\tilde{x}$ with probability 0.9 and $-\tilde{x}$ with probability 0.1. When it is stored, the stored value $\tilde{z}$ equals $\tilde{y}$ with probability 0.9 and $-\tilde{y}$ with probability 0.1.

(a) If $\tilde{x} = 1$, what is the probability that $\tilde{z} = 1$?

(b) Are $\tilde{y}$ and $\tilde{z}$ conditionally independent given $\tilde{x}$?

(c) Are $\tilde{x}$ and $\tilde{z}$ conditionally independent given $\tilde{y}$?

4. (Dead fish) Two species of fish, which we call a and b, live in a river. A biologist wants to identify dead fish found in the river. She models the problem probabilistically. First, she determines that there are roughly the same number of a fish as of b fish, so the probability of a dead fish being a is the same as the probability of it being b. Second, a fishes live at the beginning of the river and b fishes at the end. She models the river as a unit interval. Then she models the position of a dead fish as a random variable $\tilde{x}$. She estimates the conditional pdfs of $\tilde{x}$ given the species $\tilde{s}$ of the fish and determines that they are equal to $2 - 2x$ (given $\tilde{s} = a$) and $2x$ (given $\tilde{s} = b$) when $0 \leq x \leq 1$, and zero otherwise. The pdfs are shown in the following figure.



   (a) Compute the pdf of $\tilde{x}$.

   (b) If a fish is found at position 0.25, what is the probability that it belongs to species b?

   (c) Compute the conditional cdf of $\tilde{x}$ given the species of the fish.

   (d) Apply inverse transform sampling to simulate a sample from the joint distribution of the species and the position of the dead fish, using the following two independent samples obtained from a uniform distribution: 0.8, 0.64.

   (e) She finds two fish. Let $\tilde{x}_1$ be the position of the first fish, and $\tilde{x}_2$ the position of the second. $\tilde{x}_1$ and $\tilde{x}_2$ are independent and each have the same distribution as $\tilde{x}$. Derive an expression for the expected value of the distance between them $(\tilde{x}_2 - \tilde{x}_1)^2$ as a function $\sigma^2_{\tilde{x}}$ of the variance of $\tilde{x}$ that is valid for any distribution of $\tilde{x}$ (that has finite variance).

5. (Interview) A company is interviewing candidates for a data-scientist position. They estimate that the probability of a candidate being well qualified is 0.25. This is modeled by a random variable $\tilde{q}$ that equals 1 with probability 0.25, and -1 with probability 0.75. Candidates are interviewed separately by two interviewers. The decision of the interviewers are modeled as two random variables $\tilde{i}_1 = \tilde{e}_1\tilde{q}$ and $\tilde{i}_2 = \tilde{e}_2\tilde{q}$, where $\tilde{e}_1$ and $\tilde{e}_2$ are random variables that model the probability that the interviewers make a mistake. They both equal 1 with probability 0.8 (no mistake) and -1 with probability 0.2 (mistake). $\tilde{e}_1$, $\tilde{e}_2$, and $\tilde{q}$ are all mutually independent.

(a) What is the probability that the outcome of both interviews is positive, i.e. that $\tilde{i}_1 = 1$ and $\tilde{i}_2 = 1$?

(b) Are $\tilde{i}_1$ and $\tilde{i}_2$ independent?

(c) Are $\tilde{i}_1$ and $\tilde{i}_2$ conditionally independent given $\tilde{q}$?

6. (Self-driving cars) A company offers rides in Manhattan using small self-driving cars that have a maximum capacity of 2 people and can also drive around empty. People can get in and out of the cars at certain stops. A data analyst models the user behavior as a time-homogeneous Markov chain. The states of the Markov chain correspond to the number of people in the car after each stop. The analyst makes the following assumptions:

- At each stop a maximum of one person gets in. This happens with probability p_{in}, unless the car is full and nobody gets out (in that case nobody gets in).
- At each stop a maximum of one person gets out. This happens with probability p_{out}, unless the car is empty (in that case nobody gets out).
- The event of a person getting in is independent from the event that a person gets out, except if there are two people in the car. In that case, a person gets in with probability p_{in} only if someone gets out at that stop (which happens with probability p_{out}).

(a) Would it make sense to model the Markov chain as time-homogeneous in practice? Why? Answer with one sentence.

(b) For a particular choice of p_{in} and p_{out} the transition matrix is of the form

$$T := \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{bmatrix}. \tag{4}$$

In this case, if the car starts out empty, what is the probability that it is full after 2 stops?

(c) Express the transition matrix of the Markov chain in terms of p_{in} and p_{out}.

(d) Set both p_{in} and p_{out} to zero. What does the state vector converge to? Explain briefly.

7. (Potatoes)

Your aunt in Idaho asks you to analyze the production of her potato farm. Using data from 45 years, you determine that the yearly production depends mainly on two factors: the weather and the presence of a beetle (an insect which ruins the plants). You model the weather using a random variable $\tilde{w}$ and the presence of the beetle using a random variable $\tilde{b}$. $\tilde{w} = 1$ means good weather and $\tilde{w} = 0$ means bad weather. $\tilde{b} = 1$ means that the beetle is present and $\tilde{b} = 0$ that it is absent. Out of the 45 years, the following table shows how many years had good/bad weather and in how many the beetle was present.

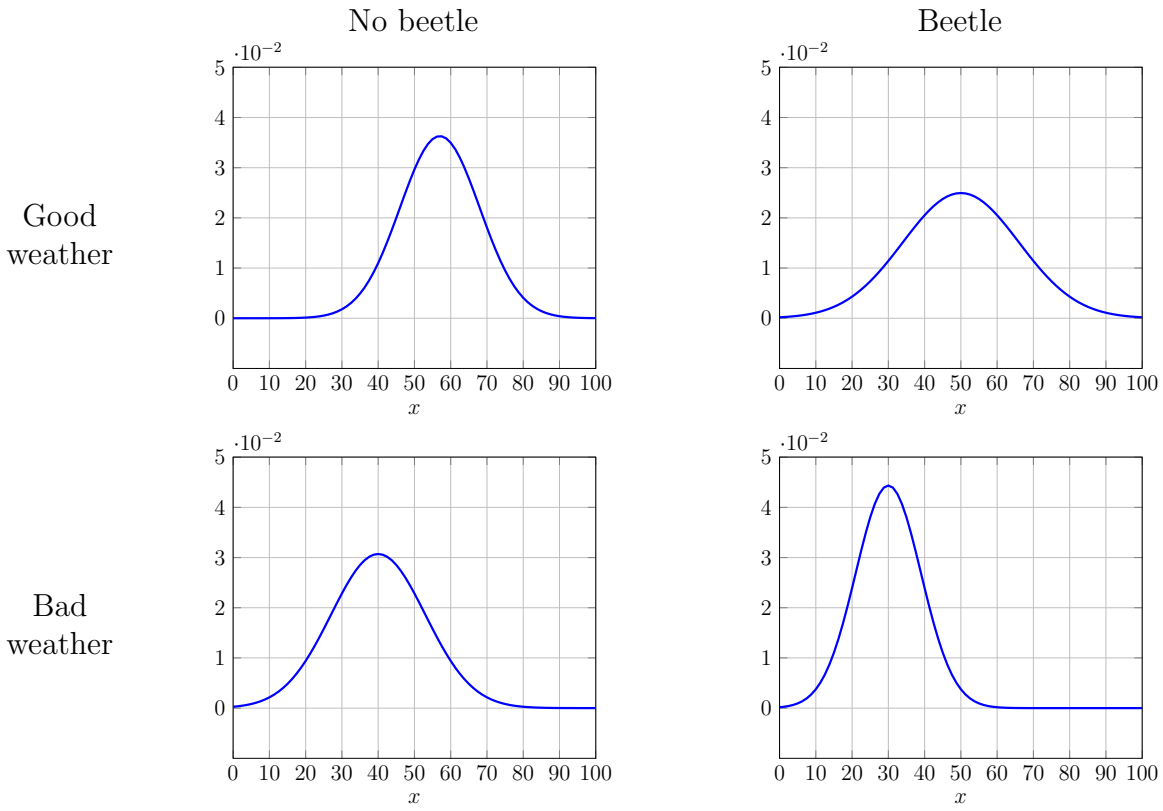


Figure 1: Estimated conditional pdf of the potato production in tons given the weather and the presence of the beetles.

	No beetle	Beetle
Good weather	5	10
Bad weather	10	20

You model the potato production during one year as a continuous random variable $\tilde{x}$. Figure 1 shows the estimated conditional pdf of $\tilde{x}$ given $\tilde{b}$ and $\tilde{w}$, obtained by fitting a Gaussian distribution to the data.

(a) What would have been the problem of using kernel density estimation with a narrow kernel width to fit a nonparametric model to the data? What would be advantage of this alternative approach if you had more data?

(b) Estimate the marginal pmfs of $\tilde{w}$ and $\tilde{b}$.

(c) According to your model, are $\tilde{w}$ and $\tilde{b}$ independent?

(d) This year the weather is good, but the beetle is present. What is the best estimate for the potato production in terms of mean square error according to your model?

(e) 50 years ago the potato production was 40 tons. There is no data regarding the beetles, but checking online you determine that the weather that year was good. Estimate of $\tilde{b}$ given this information.

(f) Are $\tilde{b}$ and $\tilde{w}$ independent given $\tilde{x}$? Justify your answer intuitively.

8. (Chad) You find your coworker Chad really annoying. He often works from home, but when he is in the office and you walk by his desk he insists on showing you pictures of his pet iguana. You would like to be able to predict when he is in the office in order to avoid him as much as possible. Another of Chad's annoying habits is to crank up the AC, so you decide to use the temperature in the office to predict his presence. After a month you have gathered the following data.

Chad	61	65	59	61	61	65	61	63	63	59
No Chad	68	70	68	64	64	-	-	-	-	-

Temperature (^{o} F)

(a) You model the temperature using a random variable $\tilde{t}$. Use a kernel density estimate which is rectangular and has width 2 to estimate the conditional pdf of $\tilde{t}$ given the presence or absence of Chad. Sketch the distribution.

(b) We model the presence or not of Chad as a parameter c ($c = 1$ means he is present, $c = 0$ that he is absent). If the temperature is 68°, does a maximum-likelihood estimate of c predict that Chad is in the office?

(c) Now we take a Bayesian approach and model the presence or absence of Chad using a random variable $\tilde{c}$ which is equal to 1 if he is there and 0 if he is not. Estimate the pmf of $\tilde{c}$ from the data.

(d) If the temperature is 64°, use the posterior distribution of $\tilde{c}$ to predict whether Chad is in the office.

(e) What problem do we run into if the temperature is 57°? Explain how using parametric estimation may alleviate this problem.

9. (Rufus) Nora's dog Rufus lives in her garden, which is the shaded area in Figure 3. After

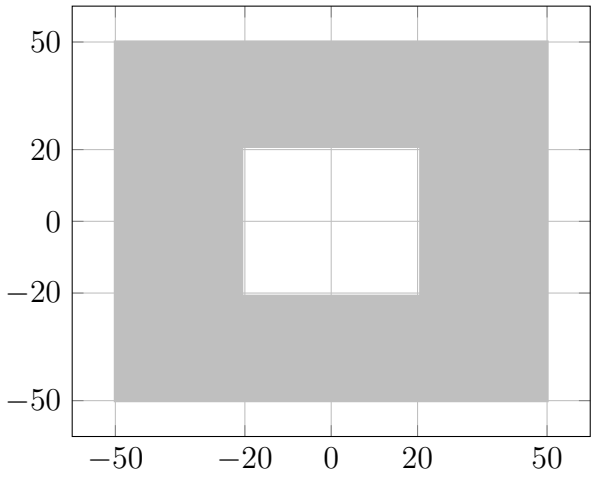


Figure 2: Nora's garden (in gray).

observing Rufus for a while she decides that his position within the garden is uniformly distributed (i.e. the probability density of his position is the same at every point of the garden).

(a) If $\tilde{x}$ is the position of Rufus on the x axis in Figure 3 and $\tilde{y}$ his position on the y axis, what is the value of the joint pdf of $\tilde{x}$ and $\tilde{y}$ on all of the x-y plane?

(b) Compute the mean and standard deviation of $\tilde{x}$.

(c) Compute the pdf of $\tilde{y}$. Sketch it.

(d) What is the pdf of $\tilde{x}$ conditioned on $\tilde{y}$? Sketch it.

(e) Are $\tilde{x}$ and $\tilde{y}$ independent? Justify your answer.

(f) Are $\tilde{x}$ and $\tilde{y}$ uncorrelated? Justify your answer. (Hint: Use iterated expectation to make your life easier).

10. (Frog) A frog lives in a garden where there are two ponds, see Figure 3. It spends 1/4 of its time in the large pond and the rest in the small pond. When it is in either of the ponds, we model its position as uniformly distributed.

(a) What is the joint pdf of the vector that indicates the position of the frog in the diagram?

(b) What is the marginal pdf of the horizontal position of the frog (i.e. its position on the horizontal axis)? Sketch the pdf.

(c) If we know that the horizontal position of the frog is 3, what is the conditional pdf of its vertical position given this information? Sketch the pdf.

(d) Is the vertical position of the frog independent from the horizontal position of the frog? Justify your answer mathematically.

(e) Is the vertical position of the frog conditionally independent from the horizontal position given the event *the frog is in the small pond*? Justify your answer mathematically.

(f) Compute the mean vertical position of the frog.

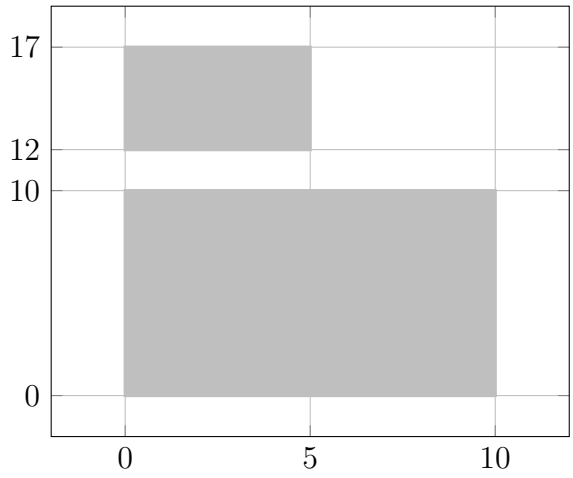


Figure 3: The two ponds.

11. (Babysitter) A babysitter is taking care of a baby. She gives him some food and then puts him to sleep. We make the following assumptions:

- The probability that the food is bad is 0.1.
- If a baby eats food that is bad, they will wake up in the middle of the night. If the food is not bad, they may still wake up (with a probability that depends on whether they are good or bad sleepers).
- All babies can be classified into *good sleepers* or *bad sleepers*. The probability that a baby that is a *good sleeper* wakes up in the middle of a given night is 0.1. The probability for a baby that is a *bad sleeper* is 0.8.
- A baby is a *good sleeper* with probability 0.6 (independently from the food).

We model the problem by defining Bernoulli random variables $\tilde{b}$ indicating whether the baby is a good ($\tilde{b} = 1$) or bad sleeper ($\tilde{b} = 0$), $\tilde{w}$ indicating whether the baby wakes up in the middle of the night ($\tilde{w} = 1$) or not ($\tilde{w}_i = 0$), and $\tilde{x}$ indicating whether the food is bad ($\tilde{x} = 1$) or not ($\tilde{x} = 0$).

(a) What is the probability that the baby wakes up in the middle of the night?

(b) If the baby wakes up in the middle of the night, what is the probability that the food was bad?

(c) Compute the probability that the food is bad conditioned on the baby waking up and being a good sleeper. Are $\tilde{b}$ and $\tilde{x}$ conditionally independent given $\tilde{w}$? Justify your answer mathematically and explain it intuitively.

12. (Sonar) A scientist is trying to determine the depth of the sea at a certain location. She knows that it must be deeper than 5 km but that is all she knows. To capture this uncertainty she models the depth as uniformly distributed between 5 and 10 km. In order to measure the depth she uses sonar, taking 2 measurements, which we model as two random variables $\tilde{s}_1$ and $\tilde{s}_2$. If the depth is equal to x then each sonar measurement is uniformly distributed between $x - 0.25$ and $x + 0.25$. The two measurements are conditionally independent given the depth.

(a) Compute and sketch the pdf of the first sonar measurement $\tilde{s}_1$.

(b) Compute the conditional pdf of the depth conditioned on the measurements being equal to 7 km and 7.1 km.

(c) Compute the joint pdf of the two sonar measurements $\tilde{s}_1$ and $\tilde{s}_2$. Are the two measurements independent? Justify your answer mathematically and explain it intuitively.

(d) If we use the average of the two measurements to estimate the depth, what is the mean squared error? (Hint: Conditioned on $\tilde{x} = x$, you can express the measurements as $\tilde{s}_1 = x + \tilde{u}_1$ and $\tilde{s}_2 = x + \tilde{u}_2$, where $\tilde{u}_1$ and $\tilde{u}_2$ are uniform random variables in $[-0.25, 0.25]$, which are jointly independent of each other and of $\tilde{x}$.)