Assignment 1: Probability fundamentals

In the following exercises, we will review some basic probability concepts. We will be using Bayes' formula, the law of total probability, the law of conditional probability, and the definition of independence. Instead of introducing these concepts here, I want to refer you to the following great resource:

- Data analysis recipes: Probability calculus for inference by David Hogg

Please read the above paper and then answer the exercises posted in it. We will not tackle all of the exercises in Hogg (2012) but focus on a selection. The following questions should be answered always in the cells below the exercise questions.

Exercise 1 (4 points)

Exercise 1.1: You have conditional pdfs $p(a \mid d), p(b \mid a, d)$, and $p(c \mid a, b, d)$. Write expressions for $p(a, b \mid d)$, $p(b \mid d)$, and $p(a \mid c, d)$.

Exercise 1.2: You have conditional pdfs $p(a \mid b, c)$ and $p(a \mid c)$ expressed or computable for any values of a, b, and c. You are not permitted to multiply these together, of course. But can you use them to construct the conditional pdf $p(b \mid a, c)$ or $p(b \mid c)$? Did you have to make any assumptions?

Answer to exercise 1

Exercise 1.1:

$p(a, b \mid d) = p(a \mid d) \cdot p(b \mid a, d)$

$p(b \mid d) = \int p(a \mid d) \cdot p(b \mid a, d) \, \mathrm{d}a$

$p(a \mid c, d) = \frac{p(c|a,d)p(a|d)}{p(c|d)} = \frac{p(c|a,d)p(a|d)}{\int p(c|a,d) \cdot p(a|d) \, \mathrm{d}a} = \frac{p(c|a,d)p(a|d)}{\int\int p(c|a,b,d) \cdot p(b|a,d) \, \mathrm{d}b \cdot p(a|d) \, \mathrm{d}a} =$

$= \frac{\int p(c|a,b,d) \cdot p(b|a,d) \, \mathrm{d}b \cdot p(a|d)}{\int\int p(c|a,b,d) \cdot p(b|a,d) \, \mathrm{d}b \cdot p(a|d) \, \mathrm{d}a}$

Exercise 1.2:

We only have the equations $p(b|a,c) = \frac{p(a|b,c)p(b|c)}{p(a|c)}$ and $p(b|c) = \int p(b|a,c) \cdot p(a|c) \, \mathrm{d}a$, hence we can only calculate the asked pdfs, if we can assume that we have one of the two.

If we assume that a and b are independent given c, then the pdfs are equal: $p(b \mid a, c) = p(b \mid c)$.

Otherwise we cannot infer them, because the pdfs in question have units in $\frac{1}{b}$, and the pdfs we know have units in $\frac{1}{a}$.

Exercise 2 (4 points)

Exercise 2.1: You have conditional pdfs $p(a \mid c)$ and $p(b \mid c)$ expressed or computable for any values of a, b, and c. Can you use them to construct the conditional pdf $p(a|b,c)$?

Exercise 2.2: You have a function $g(b)$ that is a function only of b. You have conditional pdfs $p(a \mid c)$ and $p(b \mid a, c)$. What is the expectation value $\mathbb{E}(g \mid c)$ for g conditional on c but not conditional on a?

Answer to Exercise 2

Exercise 2.1: If a and b are independent, then $p(a|b,c) = p(a|c)$. Otherwise if we also know $p(b|a,c)$, then $p(a|b,c) = \frac{p(b|a,c)p(a|c)}{p(b|c)}$. If we do not know it, but know $p(a,b|c)$, then $p(a|b,c) = \frac{p(a,b|c)}{p(b|c)}$. In any other cases we cannot construct $p(a|b,c)$.

Exercise 2.2:
$\mathbb{E}(g \mid c) = \int g(b) \cdot p(b|c) \, \mathrm{d}b = \int g(b) \cdot (\int p(b|a,c) \cdot p(a|c) \, \mathrm{d}a) \, \mathrm{d}b$

Exercise 3 (3 points)

Given the linear regression problem introduced in section 2 of Hoggs' paper, dicuss the following questions related to the likelihood function $\mathcal{L}$:

$$\mathcal{L} = \prod_n \mathcal{N}(y_n \,|\, ax_n + b, \sigma_n)$$

Exercise 3: Show that the likelihood for the model given in equations (28) through (34) in Hogg (2012) can be written in the form $Q\, exp(-\chi^2/2)$, where χ^2 is the standard statistic for weighted least-squares problems. On what does Q depend, and what are its dimensions?

Answer to Exercise 3

$$\mathcal{L} = \prod_n \mathcal{N}(D_n \,|\, ax_n + b, \sigma_n^2) = \prod_n \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{D_n-(ax_n+b)}{2\sigma_n^2}}$$

$$= \prod_n \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{y_n+\epsilon_n-ax_n-b}{2\sigma_n^2}} = \prod_n \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{ax_n+b+\epsilon_n-ax_n-b}{2\sigma_n^2}}$$

$$= \prod_n \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{\epsilon_n}{2\sigma_n^2}} = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\prod_n \sigma_n} e^{-\frac{1}{2}\sum_n \frac{\epsilon_n}{\sigma_n^2}} = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\prod_n \sigma_n} e^{-\frac{\chi^2}{2}}$$

Hence

$$Q = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\prod_n \sigma_n}$$

It depends on the parameters σ_n (standard deviations) and its dimensions are 1 x p, where p is the number of features in x_n.

Exercise 4 (4 points)

Exercise 4: The likelihood in equation (32) in Hogg (2012) is a product of Gaussians in D_n. At fixed data and b, what shape will it have in the a direction? That is, what functional form will

it have when thought of as being a function of a? You will have to use the properties of Gaussians (and products of Gaussians).

Answer to Exercise 4

$$p(\{D_n\}_{n=1}^N \mid \theta, I) = \prod_{n=1}^N p(D_n \mid \theta, I) = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\prod_n \sigma_n} e^{-\frac{\chi^2}{2}}$$

as seen above, hence it is a constant function of a, its shape is a straight horizontal line.

Exercise 5 (2 points)

Exercise 4: Show that if you take the model in equations (28) through (34) in Hogg (2012) and put a Gaussian prior pdf on a and an *independent* Gaussian prior pdf on b that your posterior pdf for a and b will be a two-dimensional Gaussian. Feel free to use informal or even hand-waving arguments; there are no mathematicians present.

Answer to Exercise 5

We know that $p(a) = \mathcal{N}(a \mid \mu_a, \sigma_a^2)$ and $p(b) = \mathcal{N}(b \mid \mu_b, \sigma_b^2)$. Because a and b are independent we also know that $p(a, b) = p(a)p(b)$. Using this, the equation 10 in Hogg (2012), and that a and b are independent from the data $\mathbf{x}$ we have that

$$\begin{aligned} p(a, b \mid \mathbf{x}, \mathbf{D}) &= \frac{1}{Z} p(\mathbf{D} \mid a, b, \mathbf{x}) p(a, b \mid \mathbf{x}) = \frac{1}{Z} p(\mathbf{D} \mid a, b, \mathbf{x}) p(a) p(b) \\ &= \frac{1}{Z} \prod_n \mathcal{N}(D_n \mid a x_n + b, \sigma_n^2) \mathcal{N}(a \mid \mu_a, \sigma_a^2) \mathcal{N}(b \mid \mu_b, \sigma_b^2) \\ &= \frac{1}{Z'} exp(-\frac{a - \mu_a}{2\sigma_a^2}) exp(-\frac{b - \mu_b}{2\sigma_b^2}) \end{aligned}$$

Z' is equal to $2\pi\sigma_a\sigma_b$ as this is the only way that our function is a pdf, which it must be. Hence the posterior pdf is a two-dimensional Gaussian with parameters

$\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$ and $\begin{pmatrix} \sigma_a^2, 0 \\ 0, \sigma_b^2 \end{pmatrix}$.

Exercise 6 (2 point)

Bayesian inference often involves computing high-dimensional integrals. For example the posterior pdf is given by:

$$p(\theta \mid D, I) = \frac{1}{Z}\, p(D \mid \theta, I)\, p(\theta \mid I)$$

where

$$Z = \int p(D \mid \theta, I)\, p(\theta \mid I)\, d\theta$$

In many cases, these integrals such as Z are analytically intractable, and we have to resort to numerical methods. One of the most popular methods for this is Monte Carlo integration. Monte Carlo integration is an absolute crucial notion in modern statistics and machine learning. If you have to solve an integral that you can write down the integrand as a product of a function $f(x)$ and a pdf $p(x)$, where $\int p(x)dx = 1$, you can use Monte Carlo integration to solve it. Provided that you can sample from $p(x)$ and evaluate $f(x)$ you can approximate the integral by

$$\int f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

where x_i are samples drawn from $p(x)$.

Exercise 6: Write a Python function that computes the integral of $f(x) = x^2$ over the interval $[1, 3]$ using Monte Carlo integration. Use $N = 1000$ samples to compute the integral. Compare your result with its true value.

```
The integral calculated by the Monte Carlo method is 8.526935483206897
The true integral is 8.666666666666668
```

The Fisher information matrix

The Fisher information matrix is the expectation value of the variance of the log-likelihood function, which is often just the expectation of the Hessian of the log-likelihood function. To obtain it, you need to compute the second derivatives with respect to the model parameters. For a data set D and model parameters θ, the Fisher information matrix is defined as:

$$\mathcal{F}_{jk} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_j\,\partial\theta_k}\ln\mathcal{L}(D\,|\,\theta)\right]$$

In this case, our model for the data is one dimensional and linear:

$$f(x\,;\,\theta) = a\,x + b \qquad \text{with} \quad \theta = (a, b)$$

and we will assume that the data are generated from some process such that the uncertainties σ_y are Gaussian (I'll drop the subscript y). Our likelihood function L for a single data point (x_i, y_i, σ_i) is then

$$L(y_i\,;\,\theta) = \mathcal{N}(y_i\,|\,f(x\,;\,\theta), \sigma_i)$$

In this model, the data points are independent, so the total likelihood $\mathcal{L}$ is just the product of the likelihoods for each individual data point:

$$\mathcal{L} = \prod_i L(y_i\,;\,\theta) \tag{1}$$

$$= \prod_i \mathcal{N}(y_i\,|\,f_i, \sigma_i) \tag{2}$$

where $f_i = f(x_i\,;\,\theta)$.

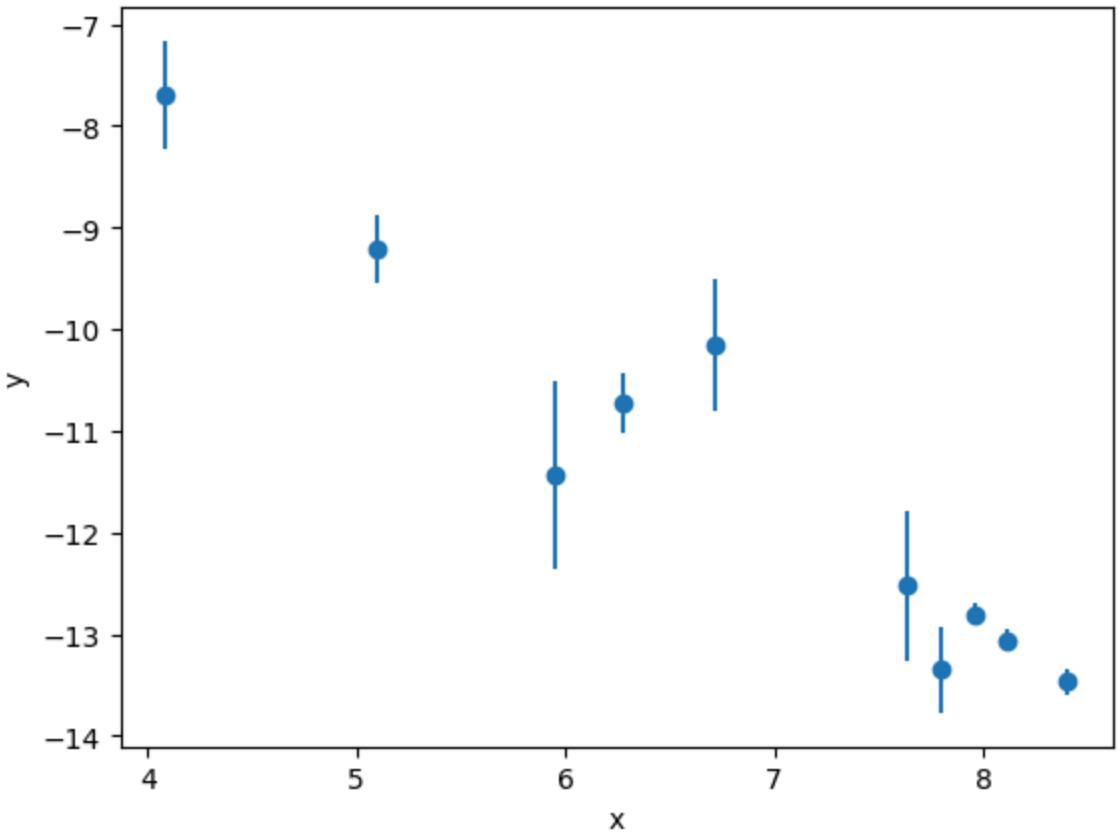
The Fisher information matrix is then the expectation value of the Hessian of the log-likelihood function. The log-likelihood function is, given Normally distributed errors:

$$\ln \mathcal{L} = -\frac{1}{2}\sum_i \left[\frac{(y_i - f_i)^2}{\sigma_i^2} + \ln(2\pi\sigma_i^2) \right]$$

Thus, we need to take second derivatives of the above expression. Although in this simple case we can do this analytically, for more complex models this expression can get increasingly difficult to write down. Thankfully, we can use automatic differentiation to do this!

Create data and set up a model for linear regression

Out[8]: [Text(0.5, 0, 'x'), Text(0, 0.5, 'y')]



Exercise 7.1 (2 points)

Exercise 7.1: Implement the likelihood function for the linear regression model described above. The likelihood function should take a dictionary of parameters `pars` and a dictionary of data `data` as arguments and return the log-likelihood of the data given the parameters. The log-likelihood should be calculated using the Gaussian likelihood function.

Hint: The jax library has a useful `jax.scipy.stats.norm.logpdf` function that you can use to calculate the log-likelihood of a single data point.

Fisher matrix via auto-diff

Having defined the likelihood function, we can now compute the Fisher matrix using automatic differentiation. The Fisher

matrix is the expectation value of the Hessian of the log-likelihood function. We can compute the Hessian of the log-likelihood function using the `jax.hessian` function. The Fisher matrix is then the negative of the Hessian of the log-likelihood function.

Exercise 7.2 (1 point)

Exercise 7.2: Interpret the values of the Fisher matrix below. What do the diagonal elements represent? What do the off-diagonal elements represent?

```
[[ 0.00566134 -0.04444363]
 [-0.04444363  0.35288954]]
```

Answer to Exercise 7.2

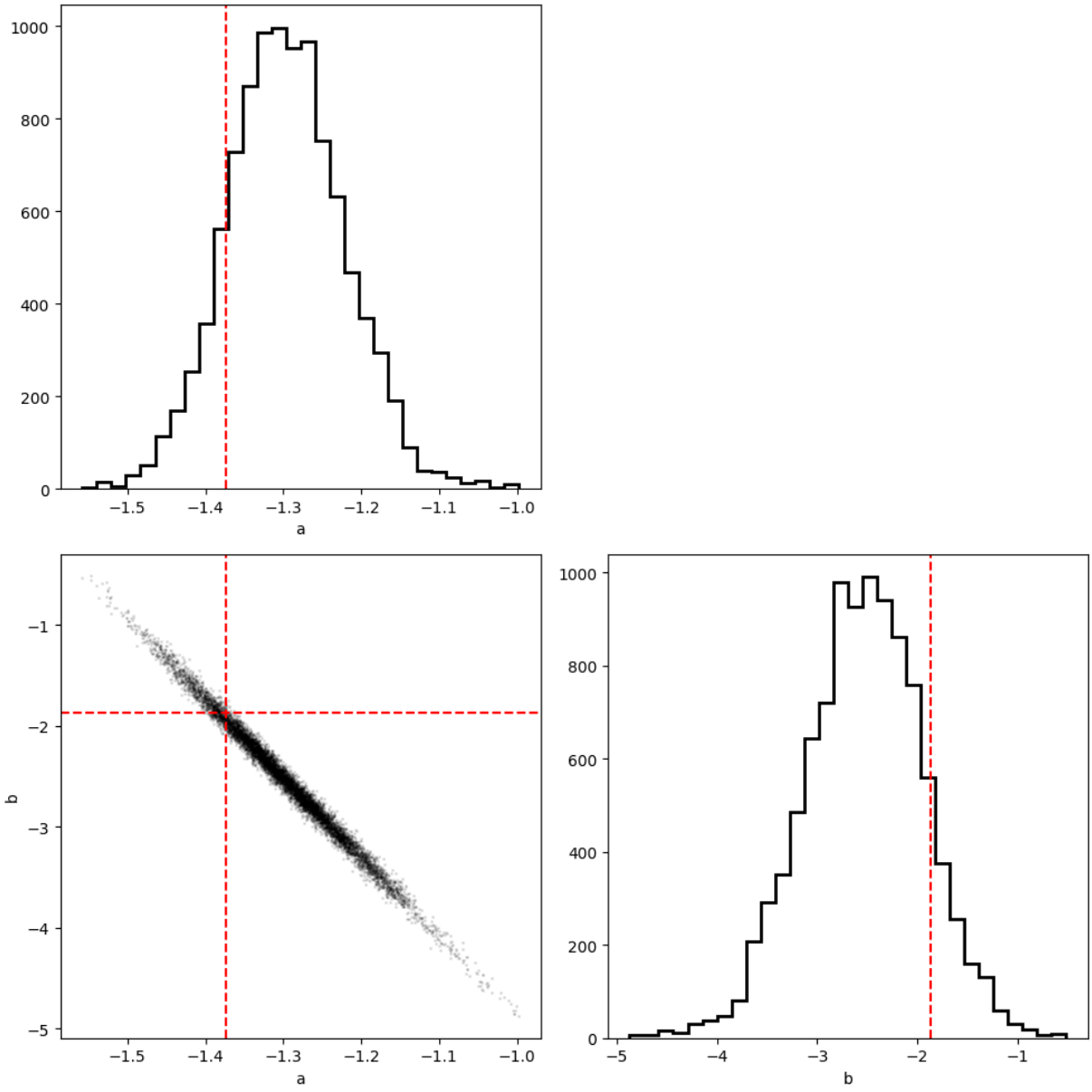
- The diagonal elements represent how much information the data provides about the individual parameters. As the entry for parameter a is way less than the one for parameter b (0.006 <<< 0.353), we can say that there is much more information about b in the data. For a we basically have no information, as the corresponding value is almost 0.
- The off-diagonal elements represenet the same thing: the correlation between the two parameters, in our case it is fairly close to 0 (-0.04), hence we can say that the parameters a and b are independent, as expected.

Exercise 8 (2 points)

Exercise 8: Define and finish the log prior and log posterior functions below to allow the application of MCMC sampling. MCMC samples the posterior distribution of the parameters a and b given the data and the likelihood function. The sampler should take the following arguments: the log posterior function, the initial parameters, the data dictionary, and the number of samples to generate.

Sample the posterior using MCMC

We will now sample the posterior using MCMC. We will use the NUTS sampler from the `blackjax` library. We will first run a warm-up phase to adapt the step size and the number of leapfrog steps. Then we will run the sampler to generate samples from the posterior distribution.



Exercise 9 (4 point)

Exercise 9.1: Interpret the results of the MCMC sampling. Compare the MCMC samples with the true parameters. Are the MCMC samples consistent with the true parameters?

Exercise 9.2: Compare the covariance matrix of the MCMC samples with the Fisher matrix. What do they both mean and are they consistent with each other?

Exercise 9.3: What are the advantages and disadvantages of using MCMC sampling compared to the Fisher matrix?

```
MCMC covariance matrix:
[[ 0.00567569 -0.04449755]
 [-0.04449755  0.35291177]]
==============================
Fisher matrix:
[[ 0.00566134 -0.04444363]
 [-0.04444363  0.35288954]]
```

Answer to Exercise 9

Exercise 9.1: The results of the MCMC sampling show normal distributions for both of our parameters, where the mean of the distribution is rather close to the true parameters. It is also worth noting that parameter a is less than the corresponding sample mean, while parameter b is above it. This is because of the way the data is created, if a decreases b has to increase as we can also see on the scatterplot.

Exercise 9.2: As already discussed the Fisher matrix tells us how much information lies in the data about the parameters and what is their covariance. The MCMC covariance matrix shows empricial covariance, and it is essentially the same what is in the Fisher matrix. It also shows the variances of the individual parameter samples. These values are also consistent with the corresponding values from the Fisher matrix, as our posterior distribution is Gaussian.

Exercise 9.3: The advantages are that it does not require a Gaussian posterior and we can add priors, and the disadvantages are that it is computationally more expensive and harder to implement.