Exercise 1

Show that if two random variables X and Y are independent, then their covariance $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ is zero. You may assume that the random variables are discrete.

i)

The most succint definiton of independence is that:

$$\mathbb{P}(X|Y) = \mathbb{P}(X) \tag{1}$$

In addition, the product rule states the following:

$$\mathbb{P}(X, Y) = \mathbb{P}(Y|X) \cdot \mathbb{P}(X) \tag{2}$$

Combining the knowledge given by this two equations, we achieve the **product rule for independent variables**, convenient for the further demonstration:

$$\mathbb{P}(X, Y) = \mathbb{P}(X) \cdot \mathbb{P}(Y) \tag{3}$$

ii)

On the other side, we develop the given definition of the covariance between X and Y, also to a more convenient (as well as very well known) form:

$$\begin{aligned} Cov[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned} \tag{4}$$

This is possible based on the linear properties of the expected values.

iii)

Now, developing $\mathbb{E}[XY]$ as a sum (or equivalently integral, for continuous variables) as given by its very definition and *taking the independence of X and Y into account*, shows the following:

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x \in X} \sum_{y \in Y} f(x) \cdot f(y) \cdot \mathbb{P}(x, y) = \sum_{x \in X} \sum_{y \in Y} f(x) \cdot f(y) \cdot \mathbb{P}(x) \cdot \mathbb{P}(y) \\ &= \sum_{x \in X} \sum_{y \in Y} f(x) \cdot \mathbb{P}(x) \cdot f(y) \cdot \mathbb{P}(y) = \sum_{x \in X} f(x) \cdot \mathbb{P}(x) \sum_{y \in Y} \cdot f(y) \cdot \mathbb{P}(y) \\ &= \mathbb{E}[X]\mathbb{E}[Y] \end{aligned} \tag{5}$$

Also possible based on the linear properties (the x-based variables can be "taken out" of the y-based sum). And since $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ holds, the following also does:

$$X \perp\!\!\!\perp Y \implies Cov[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 \tag{6}$$

□

Note that this is a one-side implication, and specifically that

$$Cov[X,Y] = 0 \not\Longrightarrow X \perp\!\!\!\perp Y \tag{7}$$

Exercise 2

Suppose you have observed N samples $x_1, ..., x_N$ drawn from a Gaussian distribution. Compute the maximum likelihood estimators for the mean and variance of the data, i.e.

$$\underset{\mu,\sigma^2}{max} \quad log \prod_{n=1}^{N} \mathcal{N}(x_n, \mu, \sigma^2)$$

where $\mathcal{N}(x_n, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

i)

Reformulating the problem gives a better intuition of what is asked: "choose the optimal μ and σ^2 so that the log-likelihood $\mathcal{L}$ becomes a maximal value:

$$\begin{aligned}
\mathcal{L} = log \prod_{n=1}^{N} \mathcal{N}(x_n, \mu, \sigma^2) &= \sum_{n=1}^{N} \{log(\mathcal{N}(x_n, \mu, \sigma^2))\} = \sum_{n=1}^{N} \{log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_n-\mu)^2}{\sigma^2}})\} \\
&= \sum_{n=1}^{N} \{log(\frac{1}{\sqrt{2\pi\sigma^2}}) + log(e^{-\frac{1}{2}\frac{(x_n-\mu)^2}{\sigma^2}})\} \\
&= N \cdot log(\frac{1}{\sqrt{2\pi\sigma^2}}) + \sum_{n=1}^{N} \{-\frac{1}{2}\frac{(x_n-\mu)^2}{\sigma^2}\} \\
&= N \cdot (-log(\sqrt{2\pi\sigma^2})) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \{(x_n-\mu)^2\} \\
&= N \cdot (-log((2\pi\sigma^2)^{\frac{1}{2}})) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \{(x_n-\mu)^2\} \\
&= -\frac{N}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \{(x_n-\mu)^2\} = \mathcal{L}
\end{aligned} \tag{8}$$

Now, in order to calculate the $\hat{\mu}$ and $\hat{\sigma}^2$ estimators, the corresponding partial derivatives of $\mathcal{L}$ can be calculated.

ii)

The partial derivative on μ is the following function:

$$\frac{\partial}{\partial \mu}\mathcal{L} = 0 - \frac{-2}{2\sigma^2}\sum_{n=1}^{N}\{x_n - \mu\} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\{x_n - \mu\} \tag{9}$$

The optimization objective is therefore a convex one, and can be solved analitically, by equalling the first derivative to zero. This can be assumed since this first derivative of μ has only one zero-crossing point, and the second derivative is a constant function:

$$0 = \frac{1}{\sigma^2}\sum_{n=1}^{N}\{x_n - \mu\} = \sum_{n=1}^{N}\{x_n - \mu\} \iff n\mu = \sum_{n=1}^{N}\{x_n\} \iff \mu = \frac{1}{N}\sum_{n=1}^{N}\{x_n\} = \overline{X} \tag{10}$$

Which means that the optimal estimator $\hat{\mu}$ is the mean value of all x samples.

iii)

In order to calculate the optimal σ^2 its corresponding partial derivative of $\mathcal{L}$ can be also calculated, and equalled to zero:

$$\frac{\partial}{\partial \sigma}\mathcal{L} = 0 = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}\{(x_n - \mu)^2\} \iff \frac{N}{\sigma} = \frac{1}{\sigma^3}\sum_{n=1}^{N}\{(x_n - \mu)^2\}$$
$$\iff N = \frac{1}{\sigma^2}\sum_{n=1}^{N}\{(x_n - \mu)^2\} \iff \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}\{(x_n - \mu)^2\}$$

so now it is also possible to calculate the $\hat{\sigma}^2$ estimator, by substituting μ with its estimator $\hat{\mu} = \overline{X}$:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}\{(x_n - \hat{\mu})^2\} \tag{11}$$

To better understand why is possible here to equal the derivative to zero, it must be regarded that, altough there is a cubic factor, we are optimizing on σ^2, which makes the zero-crossing solution unique, since $\frac{\partial}{\partial \sigma}\mathcal{L}$ is symmetrical on both axes (see Figure 1 for a better intuition on it).

iv)

Summarizing, **the maximum-likelihood estimators are**:

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}\{x_n\} = \overline{X}$$
$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}\{(x_n - \hat{\mu})^2\}$$

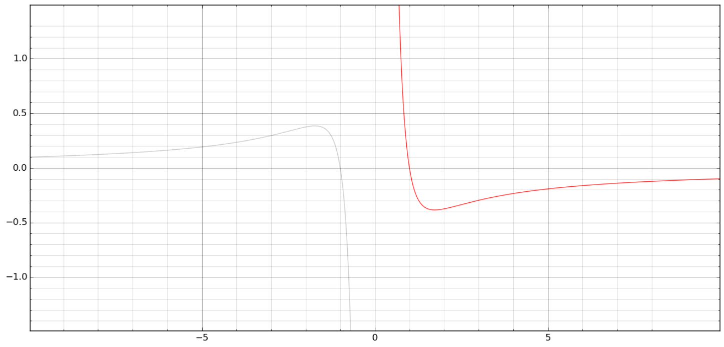

Figure 1: $f(x) = \frac{1}{x^3} - \frac{1}{x}$, similar to $\frac{\partial}{\partial \sigma}\mathcal{L}$

Exercise 3

please see/execute the python2 script "fernandez_blatt2.py" for the specifities.

i) *real parameters:* $mean_x = 2, mean_y = 4, variance_x = 4, variance_y = 2$

To see how close the estimators are to the real parameters, simply substract them to their corresponding one. It is easy to see that the more samples were produced, the more precise the estimation is.

```
estimated mean of x for 1000 samples: 2.15392196764
estimated mean of y for 1000 samples: 2.97162621077
estimated variance of x for 1000 samples: 4.10347564674
estimated variance of y for1000 samples: 2.00903592006
estimated cov(x, y) for 1000 samples: -0.021877354475

estimated mean of x for 2 samples: 3.48570375761
estimated mean of y for 2 samples: 0.855078387521
estimated variance of x for 2 samples: 0.673930874005
estimated variance of y for2 samples: 1.90147554895e-07
estimated cov(x, y) for 2 samples: -0.987359761245

estimated mean of x for 20 samples: 2.00929953147
estimated mean of y for 20 samples: 2.48505591435
estimated variance of x for 20 samples: 2.06598830411
estimated variance of y for20 samples: 1.66293241005
estimated cov(x, y) for 20 samples: -1.6741426778

estimated mean of x for 200 samples: 2.10126575922
```

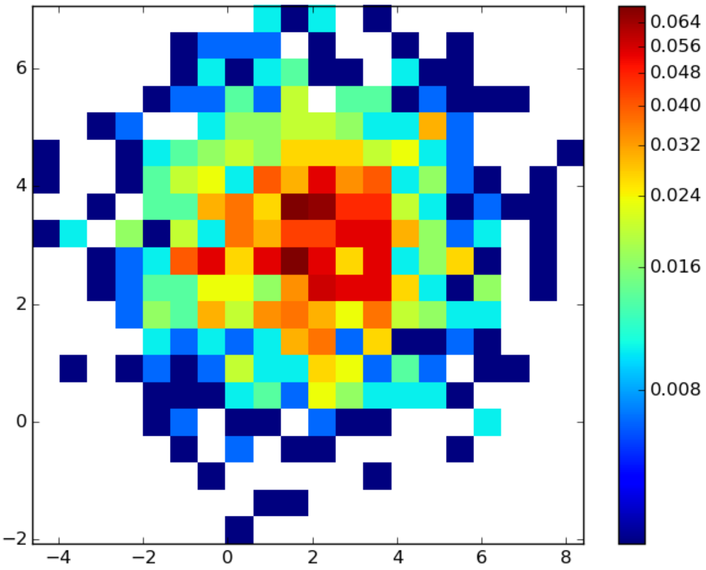

Figure 2: 2D-histogram for 20x20 bins, 1000 samples

estimat ed mean of y for 200 samples: 3.06414090786
estimated variance of x for 200 samples: 4.03243403209
estimated variance of y for200 samples: 1.78342385242
estimated cov(x, y) for 200 samples: -0.310996676007

We see also that the estimated covariance tends to zero, the more samples we use.

Exercise 4

Show that an arbitrary square matrix with elements w_{ij} can be written in the form $w_{ij} = w_{ij}^S + w_{ij}^A$ where w_{ij}^S and w_{ij}^A are symmetric and anti-symmetric matrices, respectively, satisfying $w_{ij}^S = w_{ji}^S$ and $w_{ij}^A = -w_{ji}^A$ for all i and j.

$W \in \mathbb{R}^{N\times N}$ is an arbitrary square matrix when none of its elements has to satisfy any other property than being in $\mathbb{R}$. Therefore, it suffices to show that

$$(w_{ij}^S + w_{ij}^A = w_{ij}) \text{ and } (w_{ij}^S - w_{ij}^A = w_{ji}) \quad : \quad w_{ij}^S, w_{ij}^A, w_{ij}, w_{ji} \in \mathbb{R} \tag{12}$$

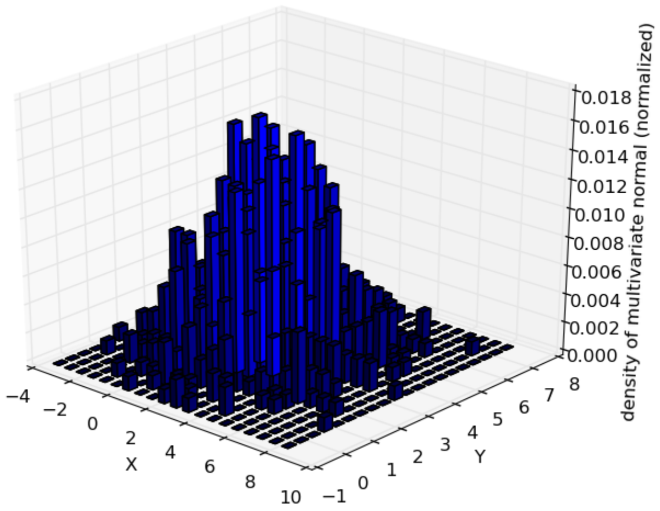

Figure 3: 3D-histogram for 20x20 bins, 1000 samples

Which can be easily done after the following transformations:

$$w_{ij}^S + w_{ij}^A = w_{ij} \iff w_{ij}^S = w_{ij} - w_{ij}^A$$
$$w_{ij}^S - w_{ij}^A = w_{ji} \iff w_{ij}^S = w_{ji} + w_{ij}^A$$
$$\iff w_{ij} - w_{ij}^A = w_{ji} + w_{ij}^A \iff 2w_{ij}^A = w_{ij} - w_{ji} \iff w_{ij}^A = \frac{w_{ij} - w_{ji}}{2}$$

□

Exercise 5

Show that a real, symmetric matrix of size $D \times D$ has $D(D+1)/2$ independent parameters.

i)

The real matrix $M \in \mathbb{R}^{D \times D}$ is symmetric when $m_{ij} = m_{ji} \quad \forall i, j \in \{1, 2, ..., D\}$. Further, independent parameters are those elements of M whose only constraint is that they have to be in $\mathbb{R}$. This holds for every element on the matrix diagonal, since every real number holds the constraint $m_{ii} = m_{ii}$. For every other element, it holds that

$$m_{ij} \text{ is independent } \iff m_{ji} \text{ is NOT independent} \qquad \forall m_{ij}, m_{ji} \in M \tag{13}$$

Simply because both of them have to be equal. The criterium to select which one of them is the independent and which the dependent is undefined. For instance, if taking (m_{ij} is independent $\iff j < i$) it can be nicely said that **every element in or above the diagonal is independent, and every other dependent**, which can be seen as the non-zero part of a triangular matrix. If counting its elements diagonally, we see that the biggest diagonal has D elements, and the diagonal size decreases by one:

$$\begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ ... & m_{22} & m_{23} & m_{24} \\ ... & ... & m_{33} & m_{34} \\ ... & ... & ... & m_{44} \end{pmatrix} \tag{14}$$

Therefore, the total number of independent elements is given by the sum

$$1 + 2 + ... + D = \sum_{d=1}^{D} \{d\} \tag{15}$$

ii)

It remains to proof that $\sum_{d=1}^{D}\{d\} = \frac{D(D+1)}{2}$, which can be done in different ways. One very intuitive way, attributed to Gauss when he was still at primary school, is to (ab)use addition's commutative property:

$$\sum_{d=1}^{D}\{d\} = 1+2+...+D = \underbrace{(1+D), (2+(D-1)), ..., (\overline{D})}_{\frac{D}{2} \text{ times}} = (D+1)\frac{D}{2} = \frac{D(D+1)}{2} \tag{16}$$

Whereas $\overline{D} = \frac{D}{2} + (\frac{D}{2} + 1) = D + 1$

□