Model answer for May 13th exercise

1. **Consider a Bernouilli random variable X defined by the parameter p: $p(X=1) = p$ and $p(X=0) = 1-p$**
    - **The Shannon entropy of a discrete random variable X that takes n possible values i for $i = 1, \dots, n$ each with probability p_i is defined as**
      $$H = -\sum_{i=1}^{n} p_i \log p_i$$
      This entropy is supposed to measure the "randomness" of a discrete random variable.
      Calculate the Shannon entropy of $b(x)$ as a function of p.
    - **What is the value of p that maximizes (resp. minimizes) the Shannon entropy for a Bernouilli distribution?**
    - **Interpret this result.**

The Shannon entropy of X is defined as

$$H(p) = -p\log p - (1-p)\log(1-p)$$

The derivative of H according to p is:

$$H'(p) = -\log p - 1 + \log(1-p) + 1 = \log\frac{1-p}{p}$$

$H'(p)$ is zero for $\frac{1-p}{p} = 1$ or $p = \frac{1}{2}$

When p is close to 0 or 1, the Shannon entropy is close to zero. Between 0 and 1 it takes its maximum value for $p = \frac{1}{2}$. So the Shannon entropy is minimal for the case of maximum certainty and maximal for the case of maximum uncertainty.

The Shannon entropy can also be defined as the expected amount of information in an event drawn from this distribution (Goodfellow et al, 2016).

2. **The Shannon entropy of a continuous random variable associated with the probability density function $p(x)$ is defined as:**
   $$H = -\int_{-\infty}^{+\infty} p(x)\log p(x)\mathrm{dx}$$
    - **This entropy is supposed to measure the "randomness" of a continuous random variable.**
      Calculate the Shannon entropy of a Gaussian distribution.
    - **When is this entropy maximum?**
    - **Interpret this result**

Consider a Gaussian random variable $N(x;\mu,\sigma^2)$.

The probability density function of x is $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

$$H = -\int_{-\infty}^{+\infty} p(x)\log p(x)\,dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left[\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 + \log\sigma\sqrt{2\pi}\right]e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx + \frac{\log\sigma\sqrt{2\pi}}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$$

The first term can easily be calculated through the change of variable $u = \left(\frac{x-\mu}{\sigma}\right)$ in the integral. It is equal to:
$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{1}{2}u^2e^{-u^2}\sigma du$, which is one half of the variance of a standardized Gaussian $N(0,1)$, that is equal to $\frac{1}{2}$.
For the second term, using the same change of variable, we obtain for the integral:

$$\int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}}\sigma du = \sigma\sqrt{2\pi}$$

So the Shannon entropy of a normal distribution is: $H = \frac{1}{2} + \log\sigma\sqrt{2\pi}.$
This expression can also be merged into a single term to obtain the best-known formula:

$$H = \frac{1}{2}\log(2\pi e\sigma^2).$$

Thus the entropy is a function of the variance (or the standard deviation) σ only, not of the mean. This is expected as the uncertainty associated with the normal distribution is controlled only by its variance (or standard deviation) only. The entropy obviously increases as σ increases.

Note that the Shannon entropy of a multivariate Gaussian function $N(\mu,\Sigma)$ in dimension n can also be calculated and is equal to:

$$H = \frac{1}{2}\log|\Sigma| + \frac{n}{2}(1+\log 2\pi)$$

For $n = 1$, we obtain $H = \frac{1}{2}\log\sigma^2 + \frac{1}{2}(1+\log 2\pi) = \frac{1}{2}\log(2\pi e\sigma^2)$.
We see that, for a fixed dimension n, what controls the randomness of a multivariate Gaussian is the determinant of its variance-covariance matrix. For instance if Σ is a diagonal matrix: $\Sigma = \begin{pmatrix}\sigma_1^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_n^2\end{pmatrix}$ we have:

$$H = \sum_{i=1}^{n}\log\sigma_i + \frac{n}{2}(1+\log 2\pi)$$

3. **Calculate the Kullback-Leibner (KL) divergence between two Gaussians $N(x;\mu_1,\sigma_1^2)$ and $N(x;\mu_2,\sigma_2^2)$.**

$$KL(f,g) = \int_{-\infty}^{+\infty} f(x)\log f(x)\,dx - \int_{-\infty}^{+\infty} f(x)\log g(x)\,dx$$

The first term is minus the already calculated entropy of $f = N(m_1, \sigma_1^2)$:

$$-\frac{1}{2} - \ \log \sigma_1\sqrt{2\pi}$$

The second term is $= -\int_{-\infty}^{+\infty} f(x)\log g(x)\,dx$

$$= \frac{1}{\sigma_1\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left[\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 + \log\sigma_2\sqrt{2\pi}\right] e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}\,dx$$

It is easy to see that the second term of this integral is equal to $\log\sigma_2\sqrt{2\pi}$.
The first term is a bit more tedious to calculate. We apply the $u = \frac{x-\mu_1}{\sigma_1}$ change of variable, then develop and we recognize one term related to the integral of the density of a $N(0,1)$, one term related to its mean (which is zero) and one term related to its variance (which is one). At the end this term is equal to:

$$\frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2}$$

Now by adding the three terms we obtain for:

$$KL(f,g) = -\frac{1}{2} - \ \log\sigma_1\sqrt{2\pi} + \log\sigma_2\sqrt{2\pi} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} = -\frac{1}{2} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} + \log\frac{\sigma_2}{\sigma_1}$$

We see that the KL divergence is zero if $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$.

Note that for two Multivariate Gaussians $N(\mu,\Sigma)$ and $N(\mu',\Sigma')$ the KL divergence is equal to :

$$KL = \frac{1}{2}\left[\log\frac{|\Sigma'|}{|\Sigma|} - n + tr(\mathbf{\Sigma}'^{-1}\mathbf{\Sigma}) + (\mu'-\mu)^T\mathbf{\Sigma}'^{-1}(\mu'-\mu)\right]$$

For instance if $\Sigma = \begin{pmatrix}\sigma_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_n\end{pmatrix}$ and $\Sigma' = \begin{pmatrix}\sigma'_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma'_n\end{pmatrix}$ are both diagonals, we have:

$$KL = \sum_{i=1}^{n}\log\frac{\sigma'_i}{\sigma_i} - n + \sum_{i=1}^{n}\frac{\sigma_i}{\sigma'_i} + \sum_{i=1}^{n}\frac{(\mu'_i-\mu_i)^2}{\sigma'^2_i}$$

4. **Suppose that the continuous random variable X follows a uniform distribution over the interval $[\mathbf{0},\mathbf{1}]$. What is the probability density function of the variable: $\boldsymbol{Y = tan\left(\pi\left(X - \frac{1}{2}\right)\right)}$?**

Let us apply the approach shown in the course this morning.
X follows the uniform distribution between 0 and 1, hence its cdf and pdf are respectively:
$F(x) = x \; and \; f(x) = 1$ for $0 \le x \le 1$.
We have, using this morning's notation, $h(x) = tan\left(\pi\left(x - \frac{1}{2}\right)\right)$
So $G(y) = P\left(tan\left(\pi\left(x - \frac{1}{2}\right) < y\right)\right) = P\left(x < \frac{1}{\pi}atan(y) + \frac{1}{2}\right) = \frac{1}{\pi}atan(y) + \frac{1}{2}$
So the pdf $g(y)$ of Y is, by differentiation:

$$g(y) = \frac{1}{\pi(1 + y^2)}$$

The approach described above allows the Monte Carlo simulation of the random variable Y which has the above pdf $g(y)$. This distribution is called the Cauchy distribution and it has the interesting property that it has no mean and no standard deviation.

5. **Suppose that z is an n-dimensional multivariate Gaussian variable: $\boldsymbol{N(z; 0, I)}$. Suppose that the n-dimensional positive definite matrix $\boldsymbol{\Sigma}$ has the Cholesky decomposition: $\boldsymbol{\Sigma = LL^T}$, where $\boldsymbol{L}$ is a lower-triangular matrix. Show that the random vector $\boldsymbol{y = \mu + Lz}$ follows a multivariate Gaussian distribution. What is its mean and what is its variance-covariance matrix**?

The y vector is multivariate Gaussian because each of its coordinates is by construction a linear combination of Gaussian random variables, hence a Gaussian random variable.
Since z is of mean zero, the mean of y will be μ and its variance covariance matrix will be:

$$E\left((y - \mu)(y - \mu)^T\right) = E\left((Lz)(Lz)^T\right) = E(Lzz^TL^T)$$
$$= LE(zz^T)L^T = LIL^T = LL^T = \Sigma$$

This provides a technique for generating Monte-Carlo simulations of a correlated multivariate Gaussian distribution.