CMPE 343: Introduction to Probability and Statistics for Computer Engineers (Fall 2021)

Homework #1

Due November 29, 2021 by 11:59pm on Moodle

Note: Please type your answers and submit your homework as PDF. To get full points, you need to show your steps clearly. If you use a theorem, rule, definition, derivation, etc. that were not covered in the lectures, you need to cite your resources. If you fail to cite your references, if you plagiarize, if you give your answers to another person, if you copy someone else's answers, your grade will be -100.

1. A function $P : A \subset \Omega \to \mathbb{R}$ is called a **probability law** over the sample space Ω if it satisfies the following three probability axioms.

   - (Nonnegativity) $P(A) \geq 0$, for every **event** A.
   - (Countable additivity) If A and B are two disjoint events, then the probability of their union satisfies
     $$P(A \cup B) = P(A) + P(B).$$
     More generally, for a countable collection of disjoint events $A_1, A_2, ...$ we have
     $$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$
   - (Normalization) The probability of the entire sample space is 1, that is, $P(\Omega) = 1$.

   (a) (5 pts) Prove, using only the axioms of probability given, that $P(A) = 1 - P(A^c)$ for any event A and probability law P where A^c denotes the complement of A.

   Proof. By the definition of A^c; A and A^c are disjoint events and also, $A \cup A^c = \Omega$.

   $$\begin{aligned} A \cup A^c &= \Omega \\ P(A \cup A^c) &= P(\Omega) \\ P(A) + P(A^c) &= P(\Omega) && \text{(Countable additivity)} \\ P(A) + P(A^c) &= 1 && \text{(Normalization)} \\ P(A) &= 1 - P(A^c) && \text{(Proof is done)} \end{aligned}$$

   □

   (b) (5 pts) Let $E_1, E_2, .., E_n$ be disjoint sets such that $\bigcup_{i_1}^{n} E_i = \Omega$ and let P be a probability law over the sample space Ω. Show that, for any event A we have
   $$P(A) = \sum_{i=1}^{n} P(A \cap E_i).$$

   Proof.

   $$\begin{aligned} A &= A \cap \Omega && \text{(by the definition of probability)} \\ &= A \cap (E_1 \cup E_2 \cup ... \cup E_n) \\ &= (A \cap E_1) \cup (A \cap E_2) \cup ... \cup (A \cap E_n) && \text{(since } \textstyle\bigcup_{i_1}^{n} E_i = \Omega\text{)} \end{aligned}$$

Thus, $P(A) = P(A \cap E_1) \cup P(A \cap E_2) \cup ... \cup P(A \cap E_n)$.
$(A \cap E_1), (A \cap E_2), ...(A \cap E_n)$ are also disjoint sets, proof of that is;

$$\begin{aligned}
&(A \cap E_1) \cap (A \cap E_2) \cap ... \cap (A \cap E_n) \\
&= A \cap (E_1 \cap E_2 \cap ... \cap E_n) && \text{(Distribution law)} \\
&= A \cap \emptyset && \text{(Since } E_1, E_2, .., E_n \text{ are disjoint sets)} \\
&= \emptyset && \text{(Null law)}
\end{aligned}$$

Hence, we can write $P(A) = P(A \cap E_1) \cup P(A \cap E_2) \cup ... \cup P(A \cap E_n)$ as;
$P(A) = P(A \cap E_1) + P(A \cap E_2) + ... + P(A \cap E_n)$.

This is equal to $P(A) = \sum_{i=1}^{n} P(A \cap E_i)$.

□

(c) (5 pts) Prove that for any two events A, B we have

$$P(A \cap B) \geq P(A) + P(B) - 1.$$

Proof. From Additive rule, we have;
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$. (1)
Also, for any event C; $0 \leq P(C) \leq 1$. So that, we can say;
$1 \geq P(A \cup B)$. (2)
If we insert (1) into (2) we find;
$1 \geq P(A) + P(B) - P(A \cap B)$. Arranging the equation, we get;
$P(A \cap B) \geq P(A) + P(B) - 1$. □

2. (10 pts) Two fair dice are thrown. Let

$$X = \begin{cases} 1, & \text{if the sum of the numbers} \leq 5 \\ 0, & \text{otherwise} \end{cases}$$

$$Y = \begin{cases} 1, & \text{if the product of the numbers is odd} \\ 0, & \text{otherwise} \end{cases}$$

What is $\text{Cov}(X, Y)$? Show your steps clearly.

Solution. We have 36 combinations for throwing 2 fair dice such as $(1,1), (1,2)...(6,6)$.
If we calculate the probability;

P(X=1, Y=1)	There are 3 occurrences for this case: (1,1), (1,3), (3,1)	3/36
P(X=1, Y= 0)	There are 7 occurrences for this case: (1,2), (1,4), (2,1), (2,2), (2,3), (3,2), (4,1)	7/36
P(X=0, Y=1)	There are 6 occurrences for this case: (1,5), (3,3), (3,5), (5,1), (5,3), (5,5)	6/36
P(X=0, Y=0)	There are 20 occurrences for this case, the rest of them	20/36

$\mu = \sum_x x \cdot f(x)$
Thus, applying the formula;
$\mu_x = 1 \times (\frac{3}{36} + \frac{7}{36}) + 0 \times (\frac{20}{36} + \frac{6}{36}) = \frac{5}{18}$

$\mu_y = 1 \times (\frac{3}{36} + \frac{6}{36}) + 0 \times (\frac{20}{36} + \frac{7}{36}) = \frac{1}{4}$

Cov(X,Y) is $\sum_x \sum_y (X - \mu_x)(Y - \mu_y) f(x, y)$. From this formula, we get;

$(1-\frac{5}{18}) \times (1-\frac{1}{4}) \times \frac{3}{36} + (1-\frac{5}{18}) \times (0-\frac{1}{4}) \times \frac{7}{36} + (0-\frac{5}{18}) \times (1-\frac{1}{4}) \times \frac{6}{36} + (0-\frac{5}{18}) \times (0-\frac{1}{4}) \times \frac{20}{36} = \frac{1}{72}$

Consequently, Cov(X,Y) = $\frac{1}{72}$. □

3. (10 pts) Derive the mean of Poisson distribution.

Solution. Poisson distribution is p(x; λt) = $\frac{e^{-\lambda t}(\lambda t)^x}{x!}$, x= 0,1,2,... Formula for mean, $\mu = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda t}(\lambda t)^x}{x!}$ (1)

$$
\begin{aligned}
\mu &= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda t}(\lambda t)^x}{x!} \\
&= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda t}(\lambda t)^x}{x \cdot (x-1)!} && \text{(opening the factorial by one element)} \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^x}{(x-1)!} && \text{(by eliminating x)} \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^x (\lambda t)^{-} (\lambda t)^1}{(x-1)!} && \text{(multiply with } (\lambda t)^{-}(\lambda t)^1\text{, which is 1)} \\
&= (\lambda t)^1 \cdot \sum_{x=0}^{\infty} \frac{e^{-\lambda t}\lambda t^{x-1}}{(x-1)!} && \text{(by arranging the multipliers)}
\end{aligned}
$$

If we look carefully, we can see that $\frac{e^{-\lambda t}\lambda t^{x-1}}{(x-1)!}$ represents the Poisson distribution for x-1.

Since the probability of the entire sample is one; $\sum_x^{\infty} \frac{e^{-\lambda t}\lambda t^{x-1}}{(x-1)!} = 1$
Thus, $(\lambda t)^1 \cdot \sum_{x=0}^{\infty} \frac{e^{-\lambda t}\lambda t^{x-1}}{(x-1)!} = \lambda t$.

Consequently, the mean of the Poisson distribution is λt. □

4. In this problem, we will explore certain properties of probability distributions and introduce new important concepts.

   (a) (5 pts) Recall Pascal's Identity for combinations: $\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}$
   Use the identity to show the following

   $$(1 + x)^N = \sum_{m=0}^{N} \binom{N}{m} \cdot x^m$$

   which is called the *binomial theorem*. *Hint*: You can use induction.
   Finally, show that the binomial distribution with parameter p is normalized, that is

   $$\sum_{m=0}^{N} \binom{N}{m} \cdot p^m \cdot (1-p)^{(N-m)} = 1$$

Proof. **Base case (N = 1):** $(1+x)^1 = \sum_{m=0}^{1} \binom{1}{m} \cdot x^m$

$$= \binom{1}{0} \cdot x^0 + \binom{1}{1} \cdot x^1 = 1 + x \text{ , It holds for N=1}$$

Inductive Hypothesis: Assume $(1+x)^k = \sum_{m=0}^{k} \binom{k}{m} \cdot x^m$ for some $k \in \mathbb{Z}$

Inductive Step: [We look for $(1+x)^{k+1} = \sum_{m=0}^{k+1} \binom{k+1}{m} \cdot x^m$, whether it holds or not.]

$(1+x)^{k+1} = (1+x) \cdot (1+x)^k$

$= (1+x) \cdot \sum_{m=0}^{k} \binom{k}{m} \cdot x^m$ (by replacing $(1+x)^k$ with our assumption)

$= (1+x) \cdot [\binom{k}{0} \cdot x^0 + \binom{k}{1} \cdot x^1 + ... + \binom{k}{k} \cdot x^m]$ (by opening the binomial)

$= \cdot [\binom{k}{0} \cdot x^0 + \binom{k}{0} \cdot x^1 + \binom{k}{1} \cdot x^1 + \binom{k}{1} \cdot x^2 + ... + \binom{k}{k} \cdot x^m + \binom{k}{k} \cdot x^{m+1}]$

(by multiplying the binomial with (1+x))

$= [1 \cdot x^0 + [\binom{k}{0} + \binom{k}{1}] \cdot x^1 + [\binom{k}{1} + \binom{k}{2}] \cdot x^2 + ... + [\binom{k}{a-1} + \binom{k}{a}] \cdot x^a + ... + 1 \cdot x^{m+1}]$

(by arranging the previous equation and knowing that $\binom{k}{0} = 1, \binom{k}{k} = 1$)

$= [[\binom{k+1}{0} \cdot x^0 + [\binom{k+1}{1} \cdot x^1 + ... + [\binom{k+1}{a} \cdot x^a + ... + [\binom{k+1}{k+1} \cdot x^{k+1}]$

(applying the Pascal's identity and $\binom{k+1}{0} = 1, \binom{k+1}{k+1} = 1$)

$= \sum_{m=0}^{k+1} \binom{k+1}{m} \cdot x^m$

So, we find that $(1+x)^{k+1} = \sum_{m=0}^{k+1} \binom{k+1}{m} \cdot x^m$ concluding that this holds for N+1. Thus binomial theorem is showed by induction.

In the binomial, there is a coefficient 1 in the each element. Meaning that;

$$(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} \cdot x^m \cdot 1^{(N-m)}$$

Proof can be done by the same way. Therefore, the binomial distribution with the parameter p is normalized is;

$$(p+1-p)^N = \sum_{m=0}^{N} \binom{N}{m} \cdot p^m \cdot (1-p)^{(N-m)}$$

Finally, this is equal to;

$$\sum_{m=0}^{N} \binom{N}{m} \cdot p^m \cdot (1-p)^{(N-m)} = 1$$

□

(b) (5 pts) Suppose you wish to transmit the value of a random variable to a receiver. In Information Theory, the average amount of information you will transmit in the process (in units of "nat") is obtained by taking the expectation of $\ln p(x)$ with respect to the distribution $p(x)$ of your random variable and is given by

$$H(x) = -\int_x p(x) \cdot \ln p(x) \cdot dx$$

This quantity is the *entropy* of your random variable. Calculate and compare the entropies of a uniform random variable $x \sim U(0,1)$ and a Gaussian random variable $z \sim \mathcal{N}(0,1)$.

Solution. For uniform random variable $x \sim U(0,1)$;

$$f(x;0,1) = \begin{cases} \frac{1}{1-0}, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Thus;

$$\begin{aligned} H(x) &= -\int_0^1 \frac{1}{1} \cdot \ln(\frac{1}{1}) \cdot dx \\ &= 0 \qquad\qquad (\text{since } \ln(1) = 0) \end{aligned}$$

For Gaussian random variable $z \sim \mathcal{N}(0,1)$;

$$\begin{aligned} f(x;0,1) &= \frac{1}{\sqrt{2\pi \cdot 1}} \cdot \exp(-\frac{(x-0)^2}{2 \cdot 1^2}), -\infty < x < \infty \\ &= \frac{1}{\sqrt{2\pi}} \cdot \exp[-\frac{x^2}{2}], -\infty < x < \infty \end{aligned}$$

Thus,

$$\begin{aligned} \ln p(x) &= \ln(\frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{x^2}{2})) \\ &= \ln(\frac{1}{\sqrt{2\pi}}) - \frac{x^2}{2} \end{aligned}$$

$$\begin{aligned} H(x) &= -\int_x p(x) \cdot \ln p(x) \cdot dx \\ &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{x^2}{2}) \cdot [\ln(\frac{1}{\sqrt{2\pi}}) - \frac{x^2}{2}] \cdot dx \\ &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{x^2}{2}) \cdot \ln(\frac{1}{\sqrt{2\pi}}) \cdot dx + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{x^2}{2}) \cdot \frac{x^2}{2} \cdot dx \end{aligned}$$

Since $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{x^2}{2}) \cdot dx$ corresponds to the probability distribution function, the integration of whole range gives 1. Therefore, integral becomes;

$$-\ln(\frac{1}{\sqrt{2\pi}}) + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{x^2}{2}) \cdot \frac{x^2}{2} \cdot dx$$

Since $\int_{-\infty}^{\infty} p(x) \cdot (x - \mu_x)^2 \cdot dx = \sigma^2$, for standard Gaussian distribution, this is equal to; $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{x^2}{2}) \cdot \frac{x^2}{2} \cdot dx = 1$. Hence, integral gives;

$$-\ln(\tfrac{1}{\sqrt{2\pi}}) + \tfrac{1}{2}$$

$$= \ln\sqrt{2\pi} + \tfrac{1}{2}$$
$$= 1.419$$

Consequently, we find that the entropy of the Gaussian random variable is larger than the entropy of the uniform random variable.

□

(c) In many applications, e.g. in Machine Learning, we wish to approximate some probability distribution using function approximators we have available, for example deep neural networks. This creates the need for a way to measure the similarity or the *distance* between two distributions. One proposed such measure is the *relative entropy* or the Kullback-Leibler divergence. Given two probability distributions p and q the KL-divergence between them is given by

$$KL(p||q) = \int_{-\infty}^{\infty} p(x) \cdot \ln\frac{p(x)}{q(x)} \cdot dx$$

i. (2 pts) Show that the KL-divergence between equal distributions is zero.
Let p(x) and q(x) be equal distributions. Then, $\frac{p(x)}{q(x)} = 1$.

Therefore, $\ln\dfrac{p(x)}{q(x)} = ln1 = 0$
Thus, KL-divergence;

$$KL(p||q) = \int_{-\infty}^{\infty} p(x) \cdot 0 \cdot dx$$
$$= 0$$

ii. (2 pts) Show that the KL-divergence is not symmetric, that is $KL(p||q) \neq KL(q||p)$ in general. You can do this by providing an example.

Assume that p(x) be a binomial distribution with n=1 and $p = 0.7$. Also, assume that q(x) be a uniform distribution with 2 possible values each having the possibility 0.5.
For binomial distribution p(x);

$$p(x) = \begin{cases} 0.3, & x = 0 \\ 0.7, & x = 1 \end{cases}$$

For uniform distribution q(x);

$$q(x) = \begin{cases} 0.5, & x = 0 \\ 0.5, & x = 1 \end{cases}$$

For $KL(p||q)$;

$$\begin{aligned} KL(p||q) &= \sum_x p(x) \cdot \ln \frac{p(x)}{q(x)} \\ &= 0.3 \times \ln \frac{0.3}{0.5} + 0.7 \times \ln \frac{0.7}{0.5} \\ &= 0.082 \end{aligned}$$

For $KL(q||p)$;

$$\begin{aligned} KL(q||p) &= \sum_x q(x) \cdot \ln \frac{q(x)}{p(x)} \\ &= 0.5 \times \ln \frac{0.5}{0.3} + 0.5 \times \ln \frac{0.5}{0.7} \\ &= 0.087 \end{aligned}$$

$$0.0082 \neq 0.0087$$

As can be seen in the example, KL-divergence is not symmetric.

iii. (16 pts) Calculate the KL divergence between $p(x) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $q(x) \sim \mathcal{N}(\mu_2, \sigma_2^2)$ for $\mu_1 = 2, \mu_2 = 1.8, \sigma_1^2 = 1.5, \sigma_2^2 = 0.2$. First, derive a closed form solution depending on $\mu_1, \mu_2, \sigma_1, \sigma_2$. Then, calculate its value. (Only numerical answer without clearly showing your steps will not be graded.)

Remark: We call this measure a *divergence* since a proper *distance* function must be symmetric.

$$KL(p||q) = \int_{-\infty}^{\infty} p(x) \cdot \ln \frac{p(x)}{q(x)} \cdot dx \tag{1}$$

$$= \int_{-\infty}^{\infty} p(x) \cdot \ln p(x) \cdot dx - \int_{\infty}^{\infty} p(x) \cdot \ln q(x) \cdot dx$$

(by the properties of the logarithmic function) (2)

The probability distribution function of Gaussian random variable is;

$$\frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot \exp[-\frac{(x-\mu)^2}{2 \cdot \sigma^2}], -\infty < x < \infty$$

$$\begin{aligned} So, \ln p(x) &= \ln(\frac{1}{\sqrt{2\pi \cdot \sigma_p^2}} \cdot \exp[-\frac{(x-\mu_p)^2}{2 \cdot \sigma_p^2}]) \\ &= \ln \frac{1}{\sqrt{2\pi \cdot \sigma_p^2}} + \ln \exp[-\frac{(x-\mu_p)^2}{2 \cdot \sigma_p^2}]) \\ &= -\frac{1}{2} \cdot \ln(2\pi\sigma_p^2) - \frac{(x-\mu_p)^2}{2 \cdot \sigma_p^2} \end{aligned}$$

From this, we can write $\ln q(x)$ as well;

$$\ln q(x) = -\frac{1}{2} \cdot \ln(2\pi\sigma_q^2) - \frac{(x-\mu_q)^2}{2\cdot\sigma_q^2}$$

If we insert this $\ln p(x)$ and $\ln q(x)$ into (2), we get;
$\int_{-\infty}^{\infty} p(x) \cdot [-\frac{1}{2} \cdot \ln(2\pi\sigma_p^2) - \frac{(x-\mu_p)^2}{2\cdot\sigma_p^2}] \cdot dx - \int_{-\infty}^{\infty} p(x) \cdot [-\frac{1}{2} \cdot \ln(2\pi\sigma_q^2) - \frac{(x-\mu_q)^2}{2\cdot\sigma_q^2}] \cdot dx$
Dividing this integral into parts;

$= -\frac{1}{2} \cdot \ln(2\pi\sigma_p^2) \int_{-\infty}^{\infty} p(x) \cdot dx - \frac{1}{2\cdot\sigma_p^2} \int_{-\infty}^{\infty} p(x) \cdot (x-\mu_p)^2 \cdot dx + \frac{1}{2} \cdot \ln(2\pi\sigma_q^2) \int_{-\infty}^{\infty} p(x) \cdot dx + \frac{1}{2\cdot\sigma_q^2} \int_{-\infty}^{\infty} p(x) \cdot (x-\mu_q)^2 \cdot dx$

$= -\frac{1}{2} \cdot \ln(2\pi\sigma_p^2) \int_{-\infty}^{\infty} p(x) \cdot dx - \frac{1}{2\cdot\sigma_p^2} \int_{-\infty}^{\infty} p(x) \cdot (x-\mu_p)^2 \cdot dx + \frac{1}{2} \cdot \ln(2\pi\sigma_q^2) \int_{-\infty}^{\infty} p(x) \cdot dx + \frac{\int_{-\infty}^{\infty} p(x) \cdot (x^2 - 2x\mu_q + \mu_q^2) \cdot dx}{2\cdot\sigma_q^2}$

$= -\frac{1}{2} \cdot \ln(2\pi\sigma_p^2) \int_{-\infty}^{\infty} p(x) \cdot dx - \frac{1}{2\cdot\sigma_p^2} \int_{-\infty}^{\infty} p(x) \cdot (x-\mu_p)^2 \cdot dx + \frac{1}{2} \cdot \ln(2\pi\sigma_q^2) \int_{-\infty}^{\infty} p(x) \cdot dx + \frac{\int_{-\infty}^{\infty} p(x) \cdot x^2 \cdot dx - \int_{-\infty}^{\infty} p(x) \cdot 2x\mu_q \cdot dx + \int_{-\infty}^{\infty} p(x) \cdot \mu_q^2 \cdot dx}{2\cdot\sigma_q^2}$

We can immediately say that $\int_{-\infty}^{\infty} p(x) \cdot dx = 1$ since the integral of whole range for a probability distribution function is 1. Also we have,
(1) $\int_{-\infty}^{\infty} x^2 \cdot p(x) \cdot dx = \mathrm{E}\langle x^2 \rangle$
(2) $\int_{-\infty}^{\infty} \mu_q \cdot 2x \cdot p(x) \cdot dx = 2 \cdot \mathrm{E}\langle x \rangle \cdot \mu_q$
(3) $\frac{1}{2\cdot\sigma_p^2} \int_{-\infty}^{\infty} p(x) \cdot (x-\mu_p)^2 \cdot dx = \frac{1}{2}$, since $(x-\mu_p)^2 = \sigma_p^2$.

Therefore, the integral gives;
$$-\frac{1}{2} \cdot \ln(2\pi\sigma_p^2) - \frac{1}{2} + \frac{1}{2} \cdot \ln(2\pi\sigma_q^2) + \frac{\mathrm{E}\langle x^2 \rangle - 2\cdot\mathrm{E}\langle x \rangle \cdot \mu_q + \mu_q^2}{2\cdot\sigma_q^2}$$

From the alternative formula for the variance, we have $\sigma_x^2 = \mathrm{E}\langle x^2 \rangle - \mu_x^2$. Thus, $\mathrm{E}\langle x^2 \rangle = \sigma_p^2 + \mu_p^2$. Also, $\mathrm{E}\langle x \rangle = \mu_p$.

Therefore,
$$-\frac{1}{2} \cdot \ln(2\pi\sigma_p^2) - \frac{1}{2} + \frac{1}{2} \cdot \ln(2\pi\sigma_q^2) + \frac{\sigma_p^2 + \mu_p^2 - 2\mu_p\mu_q + \mu_q^2}{2\cdot\sigma_q^2}$$
$$= -\frac{1}{2} \cdot \ln(2\pi\sigma_p^2) - \frac{1}{2} + \frac{1}{2} \cdot \ln(2\pi\sigma_q^2) + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\cdot\sigma_q^2}$$
Finally, this equation simplifies to;
$$= \ln(\frac{\sigma_q}{\sigma_p}) + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\cdot\sigma_q^2} - \frac{1}{2}$$

Hence, KL divergence between $p(x) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $q(x) \sim \mathcal{N}(\mu_2, \sigma_2^2)$ is given by the formula $\ln(\frac{\sigma_q}{\sigma_p}) + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\cdot\sigma_q^2} - \frac{1}{2}$.

If we insert the given values;
$$\ln(\frac{\sqrt{0.2}}{\sqrt{1.5}}) + \frac{1.5 + (2-1.8)^2}{2\cdot 0.2} - \frac{1}{2}$$
$$= 2.34$$

5. In this problem, we will explore some properties of random variables and in particular that of the Gaussian random variable.

   (a) (7 pts) The convolution of two functions f and g is defined as
   $$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau$$

One can calculate the probability density function of the random variable $Z = X+Y$ using convolution operation with X and Y independent and continuous random variables. In fact,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(\tau) f_Y(z-\tau) d\tau$$

Using this fact, find the probability density function of $Z = X + Y$, where X and Y are independent standard Gaussian random variables. Find μ_Z, σ_Z. Which distribution does Z belong to? (Hint: use $\sqrt{\pi} = \int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x$)

Solution. $f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\cdot\sigma_x^2}} \cdot \exp[-\frac{(\tau-\mu_x)^2}{2\cdot\sigma_x^2}] \frac{1}{\sqrt{2\pi\cdot\sigma_y^2}} \cdot \exp[-\frac{(z-\tau-\mu_y)^2}{2\cdot\sigma_y^2}] d\tau$

$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \exp[-\frac{(\tau-\mu_x)^2}{2\cdot\sigma_x^2}] \cdot \exp[-\frac{(z-\tau-\mu_y)^2}{2\cdot\sigma_y^2}] d\tau$

$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \exp[-\frac{(\tau-\mu_x)^2}{2\cdot\sigma_x^2} - \frac{(z-\tau-\mu_y)^2}{2\cdot\sigma_y^2}] d\tau$

$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \exp[-\frac{\sigma_y^2\cdot(\tau-\mu_x)^2}{2\cdot\sigma_x^2\cdot\sigma_y^2} - \frac{\sigma_x^2\cdot(z-\tau-\mu_y)^2}{2\cdot\sigma_y^2\cdot\sigma_x^2}] d\tau$

$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \exp[-\frac{\sigma_y^2\cdot(\tau-\mu_x)^2}{2\cdot\sigma_x^2\cdot\sigma_y^2} - \frac{\sigma_x^2\cdot(z-\tau-\mu_y)^2}{2\cdot\sigma_y^2\cdot\sigma_x^2}] d\tau$

$= \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \int_{-\infty}^{\infty} \exp[-\frac{\sigma_y^2\cdot(\tau-\mu_x)^2}{2\cdot\sigma_x^2\cdot\sigma_y^2} - \frac{\sigma_x^2\cdot(z-\tau-\mu_y)^2}{2\cdot\sigma_y^2\cdot\sigma_x^2}] d\tau$

By changing variables $t = \tau - \mu_x$, $dt = d\tau$. Also, $(z - \mu_x - \mu_y) = u$ for simplification, integral becomes;

$= \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \int_{-\infty}^{\infty} \exp[-\frac{\sigma_y^2\cdot(t)^2}{2\cdot\sigma_x^2\cdot\sigma_y^2} - \frac{\sigma_x^2\cdot(u-t)^2}{2\cdot\sigma_y^2\cdot\sigma_x^2}] dt$

For simplification, $a = \frac{\sigma_y^2}{2\cdot\sigma_x^2\cdot\sigma_y^2}$ and $b = \frac{\sigma_x^2}{2\cdot\sigma_y^2\cdot\sigma_x^2}$. Then, integral becomes;

$= \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \int_{-\infty}^{\infty} \exp[-a\cdot(t)^2 - b\cdot(u-t)^2] dt$

$= \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \int_{-\infty}^{\infty} \exp[-a\cdot(t)^2 - b\cdot(u-t)^2] dt$

$= \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \int_{-\infty}^{\infty} \exp[-at^2 - bu^2 + 2but - bt^2] dt$

$= \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \exp[\frac{-abu^2}{a+b}] \cdot \int_{-\infty}^{\infty} \exp[-(a+b)(t-\frac{au}{a+b})^2] dt$

With change of variables, $v = (t - \frac{au}{a+b}), dt = dv$;

$= \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \exp[\frac{-abu^2}{a+b}] \cdot \int_{-\infty}^{\infty} \exp[-(a+b)(v)^2] dv$

Using the hint;

$= \frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \sqrt{\frac{\pi}{a+b}} \cdot \exp[\frac{-abu^2}{a+b}]$

If we replace our simplifications with the real values, we find;

$\frac{1}{\sqrt{2\pi\cdot\sigma_x^2\cdot 2\pi\cdot\sigma_y^2}} \cdot \sqrt{\frac{\pi}{\frac{\sigma_y^2}{2\cdot\sigma_x^2\cdot\sigma_y^2} + \frac{\sigma_x^2}{2\cdot\sigma_y^2\cdot\sigma_x^2}}} \cdot \exp(\frac{-(\frac{\sigma_y^2}{2\cdot\sigma_x^2\cdot\sigma_y^2})(\frac{\sigma_x^2}{2\cdot\sigma_y^2\cdot\sigma_x^2})(z-\mu_x-\mu_y)^2}{\frac{\sigma_y^2}{2\cdot\sigma_x^2\cdot\sigma_y^2} + \frac{\sigma_x^2}{2\cdot\sigma_y^2\cdot\sigma_x^2}})$

By doing the necessary simplifications, we get;

$\frac{1}{\sqrt{2\pi\cdot(\sigma_x^2+\sigma_y^2)}} \cdot \exp[-\frac{(z-(\mu_x+\mu_y))^2)}{2(\cdot\sigma_x^2+\sigma_y^2)}], -\infty < x < \infty$

We can observe that this is the probability distribution function of Gaussian random variable having $\mu = \mu_x + \mu_y$ and the variance $\sigma_x^2 + \sigma_y^2$. Therefore, $Z = X + Y$ random

variable belongs to the Gaussian distribution, and since X and Y are the standard Gaussian random variables; $\mu_Z = 0; \sigma_Z = \sqrt{2}$.

□

(b) (5 pts) Let X be a standard normal Gaussian random variable and Y be a discrete random variable taking values $\{-1, 1\}$ with equal probabilities. Is the random variable $Z = XY$ independent of Y? Give a formal argument(proof or counter example) justifying your answer.

Solution. According to the given information;

$$Y = \begin{cases} 0.5, & y = 1 \\ 0.5, & y = -1 \end{cases}$$

Since $Z = XY$;

$$Z = XY = \begin{cases} X, & y = 1 \\ -X, & y = -1 \end{cases}$$

The probability distribution function of standard Gaussian random variable, X is;

$$\frac{1}{\sqrt{2\pi}} \cdot \exp[-\frac{(x)^2}{2}], -\infty < x < \infty$$

The probability distribution function of standard Gaussian random variable, $-X$ is;

$$\frac{1}{\sqrt{2\pi}} \cdot \exp[-\frac{(-x)^2}{2}], -\infty < -x < \infty$$

We can observe that these functions are not different from each other since $(-x)^2 = (x)^2$. Thus, Z does not change when Y is changed. Consequently, we find that $Z = XY$ is independent of Y. □

(c) (8 pts) Let X be a non-negative random variable. Let k be a positive real number. Define the binary random variable $Y = 0$ for $X < k$ and $Y = k$ for $X \geq k$. Using the relation between X and Y, prove that $P(X \geq k) \leq \frac{E[X]}{k}$. (Hint: start with finding $E[Y]$).

Solution. The probability mass function for Y is;

$$Y = \begin{cases} 0, & X < \text{k} \\ k, & X \geq \text{k} \end{cases}$$

The mean of Y, $E[Y] = 0 \cdot P(X < k) + k \cdot P(X \geq k) = k \cdot P(X \geq k)$ (*).
Let's investigate the expected values of X and Y for two cases: (1) $X < k$ and (2) $X \geq k$. When (1) $X < k$; $Y = 0$. Since X is a non-negative random variable, $E[X] > 0$. Hence, the expected value of X is greater than expected value of Y.

When (2) $X \geq k$; $Y = k$. From this case, we can observe that the expected value of X is greater than or equal to the expected value of Y. We conclude that, $E[Y] \leq E[X]$.

Thus;

$$
\begin{aligned}
&E[Y] \leq E[X] \\
&k \cdot P(X \geq k) \leq E[X] && \text{(Replace } E[Y] \text{ with } k \cdot P(X \geq k) \text{ (*))} \\
&P(X \geq k) \leq \frac{E[X]}{k} && \text{(Divide each side by } k\text{; } (k > 0))
\end{aligned}
$$

□

6. In this problem, we will empirically observe some of the results we obtained above and also the convergence properties of certain distributions. You may use the python libraries Numpy and Matplotlib.

   (a) (5 pts) In *3.a* you have found the distribution of Z = X+ Y. Let X and Y be Gaussian random variables with $\mu_X = -1$, $\mu_Y = 3$, and $\sigma_X^2 = 1$ and $\sigma_Y^2 = 4$. Sample 100000 pairs of X and Y and plot their sum Z = X+Y as a histogram. Is the shape of Z and its apparent mean consistent with *what you have learned in the lectures*?

   Solution. The shape of the histogram looks symmetric and the apparent mean is about 2. Since I expect $\mu_z = \mu_x + \mu_y$ as I found in the 5.a question and also, knowing that the mean is a linear function ($E[g(x) \pm h(y)] = E[g(x)] \pm E[h(y)]$), seeing the mean in the vicinity of $(3 + (-1)) = 2$ is consistent with what I have learned in the lectures. Also, we learned that the variance is, $\sigma^2_{aX \pm bY} = a^2\sigma_x^2 \pm b^2\sigma_y^2$. Hence, the variance of Z is $1 + 4 = 5$.

   □

   (b) (5 pts) Let X $B(n, p)$ be a binomially distributed random variable. One can use the normal distribution as an approximation to the binomial distribution when n is large and/or p is close to 0.5. In this case, $X \approx N(np, np(1-p))$. Show how such approximation behaves by drawing 10000 samples from binomial distribution with $n = 5, 10, 20, 30, 40, 100$ and $p = 0.2, p = 0.33, 0.50$ and plotting the distributions of samples for each case as a histogram. Report for which values of n and p the distribution resembles that of a Gaussian?

   Solution. The most resembling ones that of a Gaussian are n= 100 & p= 0.5, n=100 & p=0.33, and n=100 & p=0.2 values respectively. These histograms look more symmetric compared to others. Especially, the one that has n=100 and p=0.5 is the most similar to the Gaussian. □

   (c) (5 pts) You were introduced to the concept of KL-divergence analytically. Now, you will estimate this divergence $KL(p||q)$. Where $p(x) = \mathcal{N}(0, 1)$ and $q(x) = \mathcal{N}(0, 4)$. Sample 1000 samples from a Gaussian with mean 0 and variance 1. Call them $x_1, x_2, ..., x_{1000}$. Estimate the KL divergence as

$$\frac{1}{1000} \sum_{i=1}^{1000} \ln \frac{p(x_i)}{q(x_i)}$$

where $p(x) = \mathcal{N}(0, 1)$ and $q(x) = \mathcal{N}(0, 4)$. Calculate the divergence analytically for $KL(p||q)$. Is the result consistent with your estimate?

Solution. Yes, the result is consistent with my estimate. By putting in the formula,

$$\begin{aligned}
&\ln(\frac{\sigma_q}{\sigma_p}) + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2 \cdot \sigma_q^2} - \frac{1}{2}; \\
&\ln(\frac{2}{1}) + \frac{1 + (0 - 0)^2}{2 \cdot 4} - \frac{1}{2} \\
&= 0.32
\end{aligned}$$

Also, with the estimation, I found the result approximately 0.32 with a slightly change in each try.

□