LECTURE 2

Let us consider a scenario where \( N > 1 \) random variables \( X_1, X_2, ..., X_N \) are defined on a given set \( S \) of outcomes.

\[
\begin{align*}
&\text{Ex \#1: Each RV models} & &\text{Ex \#2: Each RV conveys} \\
&\text{a measurement of} & &\text{only one piece} \\
&\text{a given event} & &\text{of information} \\
&\text{(tetrode spike sorting)} & &\text{about the event} \\
& & &\text{(spike-count pairs)}
\end{align*}
\]

* Multivariate Distributions

We are interested in characterizing the vector: \( X = [X_1, X_2, ..., X_N]^T \)

First, we need to study how these RVs vary together \( \Rightarrow \) Remember that:

- \( \mathbb{S} \{ a_i < X_i \leq b_i \} \) is an event
- \( \mathbb{S} \{ a_i < X_i \leq b_i \} \) is defined on \( S \) \( \forall i \in \{1, 2, ..., N\} \)

Hence: \( P\left(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2, ..., a_N < X_N \leq b_N\right) \) is the probability of the intersection of \( N \) events on \( S \) \( \Rightarrow \) In analogy with what we saw for a single RV, we need a mathematical function to describe the probability of just these intersections of events \( \Rightarrow \) We define the joint PDF as:

\[
f: (x_1, x_2, ..., x_N) \in \mathbb{R}^N \rightarrow \mathbb{R} \text{ such that:}
\]

\[
P\left(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2, ..., a_N < X_N \leq b_N\right) = \int_{a_1}^{b_1} \ldots \int_{a_N}^{b_N} f(x_1, x_2, ..., x_N) \, dx_1 \ldots dx_N
\]

Note: \( S = \left\{ -\infty < X_i < +\infty \right\} \) for any \( i \in \{1, 2, ..., N\} \)

and \( \int_{b_i}^{a_i} \ldots \int_{a_i}^{b_N} f(x_1, ..., x_N) \, dx_1 \ldots dx_N \), is a function of \( x_i \) only

\[
N-1
\]
Hence, we have:

\[
P(a_i < x_i \leq b_i) = P(a_i < x_i \leq b_i, -\infty < x_1 < +\infty, -\infty < x_2 < +\infty, \ldots) =
\]

\[
= \int_{a_i}^{b_i} \left[ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, x_2, \ldots, x_N) \, dx_1 \ldots dx_{i-1} \, dx_{i+1} \ldots dx_N \right] \, dx_i
\]

\[
\Rightarrow \int_{a_i}^{b_i} f_{X_i}(x_i) \, dx_i \Rightarrow \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, \ldots, x_N) \, dx_1 \ldots dx_N
\]

by definition of pdf

PDF of the RV \(X_i\):

The pdf of the RV \(X_i\) is now called "MARGINAL" pdf.

Remember that two events \(A\) and \(B\) are independent iff \(P(A \cap B) = P(A) \cdot P(B)\) \(\Rightarrow\)

Therefore, we say that \(X_1, X_2, \ldots, X_N\) are INDEPENDENT RVs iff

\[
P(a_1 < x_1 \leq b_1, a_2 < x_2 \leq b_2, \ldots, a_N < x_N \leq b_N) = \prod_{i=1}^{N} P(a_i < x_i \leq b_i)
\]

From this definition a condition on the joint pdf follows:

\[
f(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} f_{X_i}(x_i)
\]

We characterize the vector \(X\) in term of joint probability distribution because, in general, the RVs \(X_1, X_2, \ldots, X_N\) are NOT independent \(\Rightarrow\) How do we measure the amount of dependency between RVs?

We usually measure dependency per pairs of RVs. Numerous different types of measures can be considered. The most used measures are:
Covariance
\[ \text{Cov}(X, Y) = E[(X-\mu_X)(Y-\mu_Y)] \]
\[ = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x-\mu_x)(y-\mu_y) f(x,y) \, dx \, dy \]
\[ \quad \mu_x = E(X), \quad \mu_y = E(Y) \]

Note this: \( X, Y \)-independent \( \Rightarrow f(x,y) = f_x(x)f_y(y) \Rightarrow \text{Cov}(X, Y) = 0 \)
\[ X = Y \Rightarrow \text{Cov}(X, Y) = \sigma_X^2 \text{-variance} \]
However: \( \text{Cov}(X, Y) = 0 \) \( \nRightarrow \) \( X, Y \)-independent

In fact, consider the following example:
\[ Y = X^2 \]
\[ \mu_Y = E(Y) = \int_{-\infty}^{+\infty} x f_x(x) \, dx = \int_{-\infty}^{+\infty} x f_x(x) \, dx + \int_{-\infty}^{+\infty} x f_x(x) \, dx = 0 \]
\[ f_x(x) = f_x(-x) \]
\[ E(X^2) = \int_{-\infty}^{+\infty} x^2 f_x(x) \, dx = 0 \]
\[ \text{Cov}(X, Y) = E(X(Y-\mu_Y)) = E(X^3) - \mu_Y E(X) = 0 \]

Correlation
\[ \rho_{xy} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \]
(Pearson's coefficient)

It is used instead of covariance because it is standardized and is not affected by \( X \), and \( Y \)'s own variability, i.e., it only depends on the joint variation of \( X, Y \).
Moreover, it can be shown that: \(-1 \leq \rho_{xy} \leq 1\)

- The notion of "correlation" can be used to predict the value of a RV:
\[ Y \sim \text{RV to be predicted} \]
\[ X \sim \text{RV to be used to predict } Y \]
\[ g(x) = \alpha + \beta X \quad \text{prediction function} \quad (\star) \]

It can be shown that the prediction function of form (\star) that minimizes the mean square (prediction) error \( \text{err} = E \left( (Y - g(x))^2 \right) \) is given when:

\[
\alpha = \mu_Y - \beta \mu_X \\
\beta = \frac{\rho_{XY} \sigma_Y}{\sigma_X}
\]

In this case, it can be shown that:

\( \rho_{XY} = 1 - \frac{E (Y-\alpha-\beta X)^2)}{\sigma_Y^2} \quad (\star\star) \)

Formula (\star\star) indicates that:

(i) \( \text{err} = 0 \) when \( \rho_{XY} = 1 \)
(ii) \( \text{err} \) is maximized when \( \rho_{XY} = 0 \)
(iii) \( \rho_{XY} \) is a measure of linear association between \( X \) and \( Y \)

The notion of "correlation" can be used to define the shape of a joint pdf:

Two RVs (e.g., \( X \) and \( Y \)) have a BIVARIATE normal distribution:

\[
f(x,y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1-\rho_{XY}^2}} e^{-\frac{1}{2} Q(x,y)}
\]

where

\[
Q(x,y) = \frac{1}{1-\rho_{XY}^2} \left[ \frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y} \right]
\]

\[
\begin{array}{c}
\text{f}(x,y) \\
\text{contours}
\end{array}
\]

\[
\begin{array}{c}
p=0 \\
p=0 \quad p=0.75 \quad \sigma_X = \sigma_Y \quad \sigma_X > \sigma_Y
\end{array}
\]
In the bivariate normal distribution, the contours satisfy the equation:

\[ Q(x, y) = c^* \]

with one value \( c^* \) for every contour \( \Rightarrow \) Contours are ellipses with axes:

\[
y = \frac{\sigma_y}{\sigma_x} x \quad \text{and} \quad y = -\frac{\sigma_x}{\sigma_y} x
\]

The concentration of the contours around the axes depends on \( p_{xy}, \text{i.e., } \) the concentration increases as \( p_{xy} \to \pm 1 \)

Note: \( X, Y \) bivariate (a.k.a. "jointly") normal \( \Rightarrow X \sim N(\mu_x, \sigma_x^2) \quad Y \sim N(\mu_y, \sigma_y^2) \)

**The joint normality is a condition stronger than just having normal RVs**

- **Mutual Information** - In order to give a definition for \( I(X, Y) \), let us first introduce a few preliminary notions:

a) Let us assume that \( f(x) \) and \( g(x) \) are two continuous pdf such that:

\[ f, g \text{ are both defined on an interval } (a, b) \quad \text{In this case, a measure of the distance between } f \text{ and } g \text{ is:} \]

\[
D_{KL}(f, g) = \int_a^b f(x) \log \left( \frac{f(x)}{g(x)} \right) dx - \text{KULLBACK-LEIBER (KL) DIVERGENCE}
\]

b) \( D_{KL}(\cdot, \cdot) \) satisfies a few useful conditions:

- If \( X \) is a RV with \( f(x) \) as pdf \( \Rightarrow D_{KL}(f, g) = E_X \left( \log f(X) - \log g(X) \right) = E_X (\log f(X)) - E_X (\log g(X)) \)
- $D_{KL}(g; f) \neq D_{KL}(f; g)$ in general, which means that $D_{KL}(\cdot; \cdot)$ captures (nonlinear) discrepancies between $f(\cdot)$ and $g(\cdot)$

- $f(\cdot) = \arg \min_{g(\cdot)} D_{KL}(f; g)$

- If $f(x)$ and $g(x)$ are normal pdfs with same variance $\sigma^2$ and mean $\mu_f$ and $\mu_g$, respectively, then: $D_{KL}(f; g) = \frac{(\mu_f - \mu_g)^2}{\sigma^2}$

Let us assume that $X$ and $Y$ are two RVs with joint pdf $f(x, y)$ and marginal pdf $f_x(x)$ and $f_y(y)$, respectively. Then we have:

$$I(X, Y) \equiv D_{KL}(f; f_x f_y) = \int_{a}^{b} \int_{a}^{b} f(x, y) \log \left( \frac{f(x, y)}{f(x) f(y)} \right) dx \, dy$$

intervals wherein $f(\cdot, \cdot)$ is defined

Note: mutual information (MI) provides a measure of the distance of the joint distribution from independence.

Note: If we define the vector $W = [X \, Y]^T$, we have the following:

- pdf of $W$ is $f(x, y)$. In fact: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1$

and all the other features of pdfs are satisfied by $f$

- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f(x, y) \, dx \, dy \equiv E_{W}(X \cdot Y)$ by analogy with the definition of "mean" for a single RV
Therefore, we can write:

\[ I(X,Y) = E_w \left( \log f_{X,Y} - \log f_X f_Y \right) \]

Ex.: \( X, Y \) are bivariate normal
\[ \rho_{XY} = \text{correlation between } X \text{ and } Y \]

\[ \Rightarrow I(X,Y) = -\frac{1}{2} \log \left( 1 - \rho_{XY}^2 \right) \]

\[ \Rightarrow I(X,Y) = 0 \text{ when } X, Y \text{ are independent and } I(X,Y) \rightarrow +\infty \text{ as the correlation between } X \text{ and } Y \text{ increases} \]

More in general, for any two RVs \( X \) and \( Y \) whose joint pdf is \( f(x,y) > 0 \), we have:

\( X, Y \) independent \( \Rightarrow I(X,Y) = 0 \) and \( I(f(X), \tilde{g}(X)) = I(X,Y) \) for any pair of

This was not guaranteed by correlation

In case of correlation, this would be true only if \( f \) and \( \tilde{g} \) are linear

In summary, given a vector of RVs \( X = [X_1, X_2, \ldots, X_n]^T \), we can provide a multivariate probability distribution by defining the joint pdf. In addition, we can provide all the pairwise measures of dependency aggregated in one matrix:

\[ \Sigma \triangleq \begin{bmatrix} m_{11} & \cdots & m_{1N} \\ m_{21} & \cdots & \vdots \\ \vdots & \ddots & \vdots \\ m_{N1} & \cdots & m_{NN} \end{bmatrix} \Rightarrow \text{This matrix can be envisioned as the adjacency matrix of a graph whose } N \text{ nodes are the RVs} \]

\( m_{ij} \triangleq \text{measure of dependency (e.g., correlation, mutual information, etc.)} \)
\( \ast \triangleq \text{conventional value assigned to } m_{ii} \text{ for } i=1,2,\ldots,N \text{ if the chosen measure is not definite} \)
In the special case when \( m_{ij} = \text{Cov}(X_i, Y_j) \) we have:

\[
\Sigma = \begin{bmatrix}
\sigma_1^2 & \sigma_1 \sigma_2 \rho_{12} & \cdots & \sigma_1 \sigma_N \rho_{1N} \\
\sigma_2 \sigma_1 \rho_{12} & \sigma_2^2 & \cdots & \sigma_2 \sigma_N \rho_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_N \sigma_1 \rho_{1N} & \sigma_N \sigma_2 \rho_{2N} & \cdots & \sigma_N^2
\end{bmatrix}
\]

- Covariance Matrix

Note: \( \Sigma \) - symmetric

\( \forall w \in \mathbb{R}^{N \times 1}, \text{Variance} (w^T X) = E \left( (w^T X - w^T \mu)^2 \right) = w^T \Sigma w \)

where, by definition, \( \mu \) is the vector of means, i.e.:

\[
\mu = \begin{bmatrix}
E(X_1) \\
E(X_2) \\
\vdots \\
E(X_N)
\end{bmatrix}, \quad w^T X = \sum_{i=1}^{N} w_i X_i, \quad W^T \Sigma W \geq 0 \iff \forall w \in \mathbb{R}^{N \times 1} \Rightarrow \Sigma \text{ is a positive semi-definite matrix}
\]

Because variance of \( w^T X \) is the variance of a random variable \( Y \) that is the linear combination of \( R \) random variables, it must be:

\( \forall w \in \mathbb{R}^{N \times 1} \Rightarrow \Sigma \) is a positive semi-definite matrix

\* Conditional Densities and Uncertainty

Another approach to the characterization of the vector \( X = [X_1, X_2, \ldots, X_N]^T \) exploits the notion of conditional probability. Consider this:

\( A, B \) - events

\[
P(A | B) = \frac{P(A \cap B)}{P(B)}
\]

Let us assume:

\[
A = \left\{ X \leq x \right\}, \quad B = \left\{ y < Y \leq y + \Delta y \right\} \Rightarrow P(X \leq x | y < Y \leq y + \Delta y) =
\]

\( X, Y \) are RVs
\[
\begin{align*}
\text{By definition of cdf and pdf of a RV, we have:} \\
\mathbb{P}(X \leq x, y < Y \leq y + \Delta y) &= F_X(x, y < Y \leq y + \Delta y) = \int_{-\infty}^{x} f_X(w, y < Y \leq y + \Delta y) \, dw \\
\mathbb{P}(y < Y \leq y + \Delta y) &= F_Y(y + \Delta y) - F_Y(y) = \int_{y}^{y + \Delta y} f_Y(v) \, dv \\
\mathbb{P}(X \leq x, y < Y \leq y + \Delta y) &= \int_{-\infty}^{x} \int_{y}^{y + \Delta y} f(w, v) \, dw \, dv
\end{align*}
\]

where: \( f(x, y) \) - joint pdf of \( X \) and \( Y \)
\( f_X(x), f_Y(y) \) - marginal pdf of \( X \) and \( Y \), respectively

Hence, by dropping the external integral, we have:

\[
\begin{align*}
f_X(x | y < Y \leq y + \Delta y) &= \frac{\int_{y}^{y + \Delta y} f(x, v) \, dv}{\int_{y}^{y + \Delta y} f_Y(v) \, dv} \uparrow \frac{f(x, y) \, \Delta y}{f_Y(y) \, \Delta y} \\
f_X(x | y) &= \lim_{\Delta y \to 0} f_X(x | y < Y \leq y + \Delta y) = \frac{f(x, y)}{f_Y(y)} \quad \text{Density of } X \text{ given } Y = y
\end{align*}
\]

It is also written as: \( f_{X|Y}(x | y) \)

Note: \( f(x, y) = f_{X|Y}(x | y) f_Y(y) \) \( \Rightarrow \) The joint pdf is equivalent to the pdf of a compound process of first drawing \( Y \) with marginal pdf \( f_Y \), and then drawing \( X \) with conditional pdf \( f_{X|Y} \).
Also, if \( X \) and \( Y \) are independent, then: \( f_{X|Y}(x|y) = f_X(x) \).

The notion of conditional PDF can be used to study one RV as a function of the other. Specifically, let us consider:

\[
E_x(X|Y=y) = \int_{-\infty}^{+\infty} x f_{x|y}(x|y) \, dx \quad \Rightarrow \quad E_x(X|Y=y) = \eta(y) \quad \Rightarrow \quad \eta(y) \text{ is an RV by definition of mean}
\]

Therefore, we can prove that:

\[
\begin{align*}
E_Y(\eta(Y)) & = E_Y(E_x(X|Y)) = E_x(X) \\
E_Y(P(X \leq \alpha | Y)) & = P(X \leq \alpha) = F_x(\alpha)
\end{align*}
\]

These results indicate that, by conditioning \( X \) to \( Y \), we obtain an estimate that, on average, converges to the actual \( X \).

This condition often applies to experiments involving bio-sIGNALS. For instance, consider the case: an experiment is run \( N \geq 1 \) times, each time with slightly different conditions. For each experimental trial, the recorded signal is itself random (i.e., two trials run under the same conditions would still return two different signals) \( \Rightarrow \) Hence, we have:

\( Y_i \) : a measure computed out of the signal recorded in trial \( i \) \( \Rightarrow \) \( Y_i \) is a RV affected by both signal and trial variability

\[ \Rightarrow \] \( X_i = E(Y_i) \) is a RV that only depends on trial variability

Based on the stated results, we can write:

\[
\sigma^2_{X_i} \left( E_{Y_i}(Y_i|X_i) \right) = \sigma^2_{Y_i} - E_{Y_i} \left( \sigma^2_{Y_i}(Y_i|X_i) \right)
\]
Note: the function \( \eta(y) = E_x(x | y = y) \) is called the REGRESSION of \( x \) on \( y \)

and can be, in general, nonlinear. However, if \( x \) and \( y \) are two bivariate

normal RVs with \( x \sim N(\mu_x, \sigma_x^2) \), \( y \sim N(\mu_y, \sigma_y^2) \) and \( \rho_{xy} \) \( \neq \) correlation

de \( x \) and \( y \), then:

\[
\eta(y) = \mu_x + \rho_{xy} \frac{\sigma_x}{\sigma_y} (y - \mu_y)
\]

Analogously, we can define: \( \xi(x) = E_y(y | x = x) \) and have:

\[
\xi(x) = \mu_y + \rho_{yx} \frac{\sigma_y}{\sigma_x} (x - \mu_x)
\]

\[\Rightarrow \rho_{xy} = \rho_{yx} \frac{\sigma_y}{\sigma_x} = \rho_{y|x} \frac{\sigma_x}{\sigma_y} \Rightarrow \rho_{xy} = \frac{\text{sign}(\rho_{y|x})}{\sqrt{\rho_{y|x}}} \cdot \rho_{x|y}
\]

Another useful application of conditional distribution is when the uncertainty of

\( y | x \) is less than the uncertainty of \( y \). Let us use the variance as a measure of

uncertainty and let us assume:

\[
\xi(x) = E_y(y | x = x) \Rightarrow \sigma_{y|x}^2 = (1 - \rho_{xy}^2) \sigma_y^2 \Rightarrow \text{The reduction of uncertainty is } \rho_{xy}^2 \sigma_y^2
\]

As a result, \( \rho_{xy}^2 = \frac{\sigma_y^2 - \sigma_{y|x}^2}{\sigma_y^2} \) is a measure of the information about \( y \) supplied by \( x \)

Note: \( \log \sigma_{y|x} = \log \sigma_y - \left( -\frac{1}{2} \log (1 - \rho_{xy}^2) \right) \)

\[H(x) = -\int_{-\infty}^{\infty} f_x(x) \log f_x(x) \, dx \quad \text{Entropy of the RV } X
\]

(it measures the disorder in the distribution of \( X \))
\[
H(X \mid Y = y) = - \int_{-\infty}^{+\infty} f_{x \mid y}(x \mid y) \log f_{x \mid y}(x \mid y) \, dx = \eta(y)
\]

\[
H(X \mid Y) \equiv E_Y(\eta(Y)) = -\int_{-\infty}^{+\infty} f_Y(y) \left( \int_{-\infty}^{+\infty} f_{x \mid y}(x \mid y) \log f_{x \mid y}(x \mid y) \, dx \right) \, dy
\]

Because \( f(x, y) = f_{x \mid y}(x \mid y) f_Y(y) \), it is easy to show that:

\[
H(X \mid Y) = H(X) - I(X, Y) \Rightarrow \text{The MI is the average amount (over } y \text{) by which the entropy of } X \text{ decreases given the additional information } Y = y
\]

**Sequences of Random Variables**

So far, we have assumed that all the RVs in the vector \( X = [X_1, X_2, \ldots, X_N]^T \) are defined on \( S \) and used simultaneously (i.e., we have focused on the intersection of events, each event being defined on one RV in \( X \)). Consider now the following case:

- \( X \): RV on \( S \) with \( E_X(X) = \mu_X \), \( E_X((X - \mu_X)^2) = \sigma_X^2 \)

An experiment is repeated \( N \) independent times and each time a sample of \( X \) is obtained \( \Rightarrow [X(\xi_1), X(\xi_2), \ldots, X(\xi_N)]^T \)

\[\uparrow \quad \uparrow \quad \cdots \quad \uparrow\]

\( \text{these are numbers} \)

- \( X_1: (\xi_1, \xi_2, \ldots, \xi_N) \rightarrow X(\xi_1) \in R \)
- \( X_2: (\xi_1, \xi_2, \ldots, \xi_N) \rightarrow X(\xi_2) \in R \)
- \( \text{Then we have that} \) \( [X_1, X_2, \ldots, X_N]^T \text{ is a vector of RVs defined on } S^N \)
- \( X_N: (\xi_1, \xi_2, \ldots, \xi_N) \rightarrow X(\xi_N) \in R \)
Moreover, these RVs satisfy the following conditions:

- cdf of $X_i$ is: $F_X(x) = F_X(x) \forall x \in \mathbb{R}$
- $X_1, X_2, \ldots, X_N$ are independent (as so are the $N$ consecutive experiments)

$[X_1, \ldots, X_N]^T$ is called a "random sample" and RVs $X_1, \ldots, X_N$ are said "i.i.d." (independent and identically distributed)

Now, let us suppose that the distribution of $X$ (e.g., $\mu_X$ and $\sigma_X^2$) must be estimated by using the (sample) distribution of $X_1, X_2, \ldots, X_N$, which have been computed on batches on $N$ experiments (N-tuple). We can solve this problem by using the following results:

$$\hat{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \quad (it\ is\ a\ RV\ on\ \mathbb{R})$$

**Theorem**

If $(X_1, X_2, \ldots, X_N)$ are i.i.d. then we have:

$$E_{\hat{X}}(\hat{X}) = \mu_X$$
$$E_{\hat{X}}((\hat{X} - \mu_X)^2) = \frac{\sigma_X^2}{N}$$

**Note:** This theorem indicates that combining experiments in N-tuple allows to estimate the mean $\mu_X$ of $X$ with a variance that is a fraction of $\sigma_X^2$.

Moreover, if the independence is NOT satisfied (e.g., $\text{Cov}(X_i, X_j) = \rho \sigma_X^2$ with $|\rho| < 1$, $\forall i, j$), then we have:

$$E_{\hat{X}}((\hat{X} - \mu_X)^2) = \frac{\sigma_X^2}{N} + \frac{N-1}{N} \rho \sigma_X^2$$

**Definitions**

Because we consider the (sample) distributions of $X_1, X_2, \ldots, X_N$, we will have $N$ (sample) cdfs: $F_{X_1}(x), F_{X_2}(x), \ldots, F_{X_N}(x)$. We say:

$$(X_1, X_2, \ldots, X_N) \text{ converges in distribution to } X \iff \lim_{N \to \infty} F_{X_N}(x) = F_X(x) \forall x \in \mathbb{R}$$
(X_1, X_2, \ldots, X_N) converges in probability to a constant \( c \in \mathbb{R} \) such that \( P(X = c) = 1 \)

Note: notation \( X_n \xrightarrow{D} X \) means "convergence in distribution to \( X \)."

**Theorem**  If \( (X_1, X_2, \ldots, X_N) \) are i.i.d.

\( \Rightarrow \) The sequence \( \left( \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N \right) \) with

\[ \hat{X}_i = \frac{1}{i} \sum_{j=1}^{i} X_j \]

converges in probability to \( \mu_x \)

**CLT**  If \( (X_1, X_2, \ldots, X_N) \) are i.i.d.

\[ \Rightarrow \]

Denoted with:

\[ Z_i = \frac{\hat{X}_i - \mu_x}{\sigma_x} \]

we have that the sequence \( (Z_1, Z_2, \ldots, Z_N) \) converges in distribution to \( N(0,1) \)

**Theorem** \( \overline{X} = \left[ \begin{array}{c} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_m \end{array} \right] \) where

\[ \overline{X}_j = \frac{1}{N} \sum_{i=1}^{N} X_{ij} \]

and \( X_{ij} \) are RVs for any \( i \) and \( j \)

\[ \sum \triangleq \text{Cov}(\overline{X}) \quad \text{and} \quad \Sigma > 0 \ (i.e., \text{positive definite}) \]

\[ \Rightarrow \]

Let us define:

\[ Z_N(w) = \sqrt{N} \left( \sum_{i}^{N} (X_i - \overline{X}) w^T \right) \]

\( \overline{X} = \left[ \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_m \right]^T \), vector of means of \( \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_m \)

\( R \times m \times 1 \)

\( \forall w \in \mathbb{R} \times 1 \)

\( w \neq 0 \)

\( \Rightarrow \)

\( Z_N(\overline{X}) \) converges in distribution to \( N(0,1) \)

\[ \triangleq \]

References:

Textbook: ch 4 (Bayesian Estimators NOT included)

ch 6